J. Phys.: Condens. Matter **17** (2005) 5757–5769 doi:10.1088/0953-8984/17/37/012

Using ONETEP for accurate and efficient $\mathcal{O}(N)$ density functional calculations

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

[1] Physical and Theoretical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, UK
[2] Theory of Condensed Matter, Cavendish Laboratory, J J Thomson Avenue, Cambridge CB3 0HE, UK

E-mail: chris-kriton.skylaris@chem.ox.ac.uk

Received 19 July 2005, in final form 16 August 2005
Published 2 September 2005
Online at stacks.iop.org/JPhysCM/17/5757

Abstract

We present a detailed comparison between ONETEP, our linear-scaling density functional method, and the conventional pseudopotential plane wave approach in order to demonstrate its high accuracy. Further comparison with all-electron calculations shows that only the largest available Gaussian basis sets can match the accuracy of routine ONETEP calculations. Results indicate that our minimization procedure is not ill conditioned and that convergence to self-consistency is achieved efficiently. Finally, we present calculations with ONETEP, on systems of about 1000 atoms, of electronic, structural and chemical properties of a wide variety of materials such as metallic and semiconducting carbon nanotubes, crystalline silicon and a protein complex.

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

1. Introduction

The formalism of Kohn–Sham density functional theory (DFT) [1, 2] for electronic structure calculations has become established as an approach that provides a good description of electronic correlation while keeping the size of calculations tractable. Nevertheless, the computational time taken by a conventional DFT calculation increases with the cube of the number of atoms. This scaling limits the size of problems that can be tackled to a few hundred atoms at most. As a consequence, many exciting problems which lie at the interface between the microscopic and mesoscopic worlds, particularly in the fields of biophysics and nanoscience, are out of the reach of DFT calculations. Progress towards the goal of bringing the predictive power of DFT to bear on these problems can be made only by developing approaches for DFT calculations that have linear-scaling or $\mathcal{O}(N)$ instead of cubic-scaling computational cost.

[3] Author to whom any correspondence should be addressed. http://www.chem.ox.ac.uk/researchguide/ckskylaris.html

0953-8984/05/375757+13$30.00

Even though there have been numerous theoretical developments, so far linear-scaling methods have not lived up to their early promise. Linear-scaling approaches are still described as 'experimental' [3] and so far there are few examples of successful application to problems of interest in materials or biological sciences [4]. For a review see [5, 6]. Our ONETEP linear-scaling method for DFT calculations allows for the systematic control of both truncation errors and variational freedom in the basis set. For full details, including demonstration of the linear-scaling behaviour, see [7] and references therein. Here we demonstrate that ONETEP can be used to solve real problems with the same level of confidence and general applicability as conventional cubic-scaling DFT approaches.

In section 2 we begin with a brief presentation of the formalism for linear-scaling DFT on which ONETEP is based. In section 3 we compare ONETEP with conventional well established cubic-scaling methods with emphasis on the case of systematic improvement in the basis set, and hence in accuracy, and in the speed of self-consistent convergence. In section 4 we show how ONETEP can be used to explore a range of materials with thousands of atoms ranging from nanostructures to bulk solids to biomolecules. Finally, in section 5 we present our conclusions.

2. Overview of theory

Kohn–Sham DFT enables the problem of many interacting electrons in a static external potential to be mapped onto a fictitious system of non-interacting particles. Self-consistent solution of the resulting set of single-particle Schrödinger equations gives the ground-state energy and density of the original interacting problem. All the information about the ground state of the system is contained in the single-particle density matrix $\rho(\mathbf{r}, \mathbf{r}')$ which, provided there is a bandgap in the material, decays exponentially [8–12] as a function of the distance between $\mathbf{r}'$ and $\mathbf{r}$. This property can be exploited to truncate the density matrix so that the amount of information it contains increases only linearly with the number of atoms. To perform this truncation in a practical way, the density matrix is expressed as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}') \tag{1}$$

where the $\{\phi_\alpha\}$ are a set of spatially localized, non-orthogonal generalized Wannier functions (NGWFs) [13] and the matrix $\mathbf{K}$ is called the density kernel [14]. $\mathbf{K}$ can be made sparse by enforcing the condition $K^{\alpha\beta} = 0$ when $|\mathbf{R}_\alpha - \mathbf{R}_\beta| > r_{\text{cut}}$, where $\mathbf{R}_\alpha$ and $\mathbf{R}_\beta$ are the centres of the localization regions of NGWFs $\phi_\alpha(\mathbf{r})$ and $\phi_\beta(\mathbf{r})$, respectively.

ONETEP belongs to the category of methods that aim for high accuracy by optimizing the energy self-consistently not only with respect to $\mathbf{K}$ but also with respect to the NGWFs [15–20]. In ONETEP the NGWFs are expanded in a basis set of periodic cardinal sine (psinc) functions [13, 21], also known as Dirichlet or Fourier Lagrange-mesh functions [22, 23]. Each psinc function is centred on a particular point of a regular real-space grid. Figure 1 shows how this property is used to impose localization on the NGWFs within predefined spherical regions and [24] describes in detail our methodology for the computation of each term (including the Hartree potential) in the total energy with $\mathcal{O}(N)$ cost.

3. Basis set convergence

Since the computational cost of a DFT calculation increases with the size of the basis set it is important to be able to converge calculated properties to the desired accuracy using the smallest possible basis set. The most convenient way to achieve this is to improve the basis set systematically. For instance, the quality of a plane wave basis [25] is increased via a single parameter, the kinetic energy cut-off. At the other end of the spectrum are atomic orbital (AO)

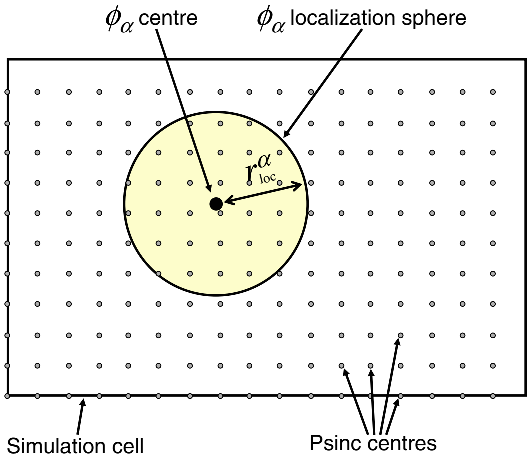

Figure 1. Imposing localization on an NGWF (ϕ_α). The NGWF is expanded only in the psinc functions whose centres fall inside its localization sphere.

basis sets which do not span space in a uniform manner and whose systematic refinement is not straightforward. An AO basis is defined by a number of independent features, such as the number of functions per atom, and their radial and angular shapes. Furthermore, unlike plane waves, AO basis sets are not orthogonal and consequently the undesired effect of linear dependence can often hinder efforts to improve their quality. Nevertheless, numerous careful attempts have been made to construct series of atomic basis sets which demonstrate systematic improvement to varying degrees [26–29]. Particular attention has been paid to Gaussian [30] functions where the series of even-tempered [31] and correlation-consistent [32] basis sets are amongst the best known cases of AO bases with systematic behaviour. In ONETEP our psinc basis is constructed from plane waves in such a way that it fully retains their desirable properties of orthogonality and systematicity whilst being localized.

It is important to note that the set of plane waves which constitute the psinc functions is different from that in a typical plane wave calculation. The relation between the two is clarified in figure 2. The psinc basis set is constructed from plane waves $e^{i\mathbf{G}\cdot\mathbf{r}}$ with wavevectors $\mathbf{G}$ belonging to a cube of side-length $2\mathbf{G}_{\mathrm{upper}}$ in reciprocal space. On the other hand, conventional

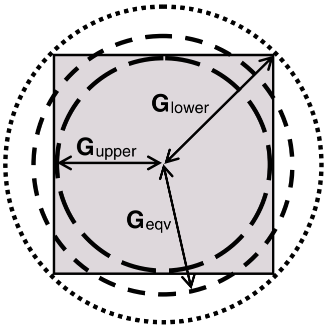
$\mathbf{G}_{\mathrm{lower}}$
$\mathbf{G}_{\mathrm{upper}}$
$\mathbf{G}_{\mathrm{eqv}}$

Figure 2. The psinc basis of ONETEP is constructed from a cube of wavevectors. Conventional plane wave approaches such as CASTEP define their plane wave basis from a sphere of wavevectors. Three choices of such spheres that could be used to compare ONETEP and CASTEP calculations are shown.

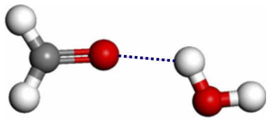

Figure 3. The molecular structure of the formaldehyde–water hydrogen bonded complex used in our tests (not equilibrium geometry).

plane wave approaches [25] such as the CASTEP code [33] construct their basis from a sphere of **G** vectors. Therefore, to compare ONETEP calculations with a code such as CASTEP, we need to decide first on the most appropriate sphere of wavevectors. Figure 2 shows some choices for the radius of this sphere: $\mathbf{G}_{\mathrm{upper}}$ (sphere inscribed in cube, the CASTEP basis is a subset of the ONETEP basis), $\mathbf{G}_{\mathrm{eqv}}$ (sphere has equal volume with cube, ONETEP and CASTEP basis sets have an equal number of functions with most of them in common) and $\mathbf{G}_{\mathrm{lower}}$ (sphere circumscribes the cube, the CASTEP basis is a superset of the ONETEP basis).

In order to examine the strengths and weaknesses of ONETEP compared to conventional plane wave and AO approaches we have carried out a series of tests on the hydrogen bond in the formaldehyde–water complex shown in figure 3. This is a rather sensitive test as hydrogen bonds are amongst the weakest and longest chemical bonds, yet they are very important because they are commonly encountered as major contributors to the structural stability and function of most biological macromolecules such as proteins, DNA and sugars [34]. For the purpose of comparison we have used the local density approximation (LDA) [35, 36] exchange–correlation (XC) functional.

Table 1 shows the binding energies we obtained from calculations with CASTEP for the three kinetic energy cut-offs in figure 2. Also shown is the total energy of the bound complex. The core electrons in these calculations were replaced by norm-conserving pseudopotentials [37–39]. Periodic boundary conditions were used and the molecule was placed in a very large cubic simulation cell (30 Å × 30 Å × 30 Å) to ensure that the supercell approximation [25] holds extremely well.

The corresponding ONETEP results are shown in table 2 for the same periodic simulation cell, pseudopotentials and LDA XC functional. A total of 16 NGWFs were used for the hydrogen bonded complex, one on each H atom and four on each C and O atom. We have performed calculations for a wide range of NGWF localization sphere radii r_{loc} and we observe that the binding energy agrees with the converged CASTEP value[4] to 1 meV, for r_{loc} as small as 3.7 Å. The total energy converges rapidly from above as a function of r_{loc}, as expected for a basis set variational method [40]. We note that, once we are converged with respect to r_{loc}, the ONETEP result lies between the CASTEP bounds shown in table 1 and, as one would expect, agrees closely with the 935 eV cut-off result ($\mathbf{G}_{\mathrm{eqv}}$ sphere in figure 2). From here on we define the psinc kinetic energy cut-off to be the kinetic energy cut-off of the plane wave sphere with the same volume as the cube of our psinc basis.

Table 2 also shows the number of self-consistency iterations taken to converge the total energy, and we make the observation that this is independent of the localization region radius r_{loc}, which demonstrates that our method does not suffer from the 'superposition ill conditioning' described in [41].

To complete our comparison we present in table 3 calculations with the AO approach as implemented in the NWChem [42] quantum chemistry program which uses Gaussian basis sets and a closely related formula [35, 50] for the LDA XC functional. In this approach the

[4] One kcal mol^{-1} is equal to 43.36 meV.

Table 1. Calculations on the formaldehyde–water complex with CASTEP, [33].

Kinetic energy cut-off (eV)	Total energy (eV/atom)	Binding energy (meV)
608 ($\propto \mathbf{G}^2_{\text{upper}}$)	−154.444	145
935 ($\propto \mathbf{G}^2_{\text{eqv}}$)	−155.044	149
1823 ($\propto \mathbf{G}^2_{\text{lower}}$)	−155.082	148

Table 2. Calculations on the formaldehyde–water complex with ONETEP [7].

r_{loc} (Å)	Number of iterations	Total energy (eV/atom)	Binding energy (meV)
2.6	13	−154.789	168
3.2	13	−154.890	155
3.7	11	−154.914	150
4.2	11	−154.921	148
4.8	12	−154.924	148

Table 3. Calculations on the formaldehyde–water complex with NWChem [42] using Gaussian basis sets of increasing size.

Basis name	Number of AOs	Binding energy (meV)	Counterpoise-corrected binding energy (meV)
STO-3G [43]	19	91	39
3-21G [44]	35	186	92
6-31G [45]	35	171	128
6-31+G* [46, 47]	65	159	143
6-31++G** [46, 47]	81	162	147
cc-pVDZ & diffuse [48, 49]	111	153	146
cc-pVTZ [48]	165	157	133
cc-pVTZ & diffuse [48, 49]	265	149	147
cc-pVQZ [48]	350	151	140
cc-pVQZ & diffuse [48, 49]	535	148	147

core electrons are treated explicitly and the molecules are virtually isolated in space as the calculations are done with open boundary conditions. The total number of AOs (contracted Gaussian functions) for the whole formaldehyde–water complex for each basis set is also shown in table 3.

From table 3 we observe that the convergence of the total energy of the complex is neither uniform nor rapid, as a consequence of the fact that the different features, e.g., diffuse functions etc, introduced to the basis set affect the energy to different extents. We also note that the size of the basis set required to reach the same level of accuracy as ONETEP is very large. The calculations with the Gaussian basis set suffer from basis set superposition error (BSSE) and thus in table 3 we also give a column with binding energies calculated with the counterpoise correction method of Boys and Bernardi [51]. This costly correction procedure significantly improves the binding energies obtained with the medium sized basis sets (6-31+G$*$ and 6-31++G$**$).

4. Nanostructures, crystals and biomolecules

In this section we present several examples of calculations on systems with around 1000 atoms. Materials and molecules with this number of atoms are usually beyond the capabilities of conventional cubic-scaling approaches.

4.1. Nanostructures: carbon nanotubes

Carbon nanotubes are at the centre of many nanotechnology applications because of their unique electronic and mechanical properties [52]. From a structural point of view nanotubes are seamless cylinders of graphene, which can be either semiconducting or metallic. A method such as ONETEP where linear scaling is achieved by taking advantage of the exponential decay of the density matrix present in insulators is not expected to be efficient on metallic systems where the decay is only algebraic [11]. Metallic systems therefore present a significant challenge and carbon nanotubes are an ideal test case that can provide us with insight into how switching from a non-metallic to a metallic system (while keeping all other factors essentially unchanged) affects calculations where density matrix truncation is applied. We have studied segments of metallic (10, 10) armchair and semiconducting (20, 0) zigzag carbon nanotubes [52, 53].

The (10, 10) nanotube segments are constructed by repeating identical units of 40 atoms while the (20, 0) segments are made of units of 80 atoms. For the (10, 10) nanotube we performed ONETEP calculations on segments consisting of 8, 15, 16, 30 and 32 units ranging from 320 to 1280 atoms. For the (20, 0) nanotube we used segments of 8 and 16 units with 640 and 1280 atoms respectively. As our nanotube segments were made of repeated identical units we were able to perform with CASTEP calculations equivalent to ONETEP by using only a single unit but equivalent Brillouin zone sampling. The same LDA [35, 36] XC functional and pseudopotential were used by both codes. The plane wave kinetic energy cut-off of CASTEP was set to 410 eV as was the psinc kinetic energy cut-off of ONETEP. In the ONETEP calculations the radii r_{loc} of the carbon NGWF localization spheres were 3.3 Å. The nanotube segments were placed in orthorhombic simulation cells with their axis aligned with the z-axis. The dimensions of the cells along the x- and y-axes were 30 Å × 30 Å. These simulation cells ensured negligible interaction of the nanotubes with their periodic images as the diameter of the (10, 10) tubes is just 13.6 Å and that of the (20, 0) tubes is 15.6 Å. In order to perform a detailed comparison of the results between the two codes we diagonalized the converged ONETEP Hamiltonian in the NGWF representation and obtained canonical molecular orbitals. From these we constructed the density of states (DOS) by smearing with Gaussians with a halfwidth of 0.1 eV. Our results are shown in figure 4. The two codes give virtually identical DOS in the important regions of 1 eV below and above the Fermi level and very close agreement in the region below −1 eV. In the region above 1 eV the agreement deteriorates rapidly. This is not surprising as the NGWFs of ONETEP are specifically optimized to describe the density matrix which is composed of occupied bands and no emphasis is placed on the description of the conduction bands. It is still remarkable that the low-lying conduction band DOS is calculated correctly with ONETEP.

As we make our (10, 10) nanotube segments longer, we increase the number of closely spaced **k**-points from the metallic band structure of the nanotube that we fold into our equivalent Γ point description of the band structure and the density matrix. We have found that as the number of segments increases it becomes more and more difficult to impose a finite density kernel cut-off threshold r_{cut} in ONETEP while maintaining any degree of accuracy. With the 30- and 32-unit segments an infinite r_{cut} becomes essential in order to obtain useful results. In contrast, the (20, 0) nanotube remains amenable to density kernel truncation as the length of

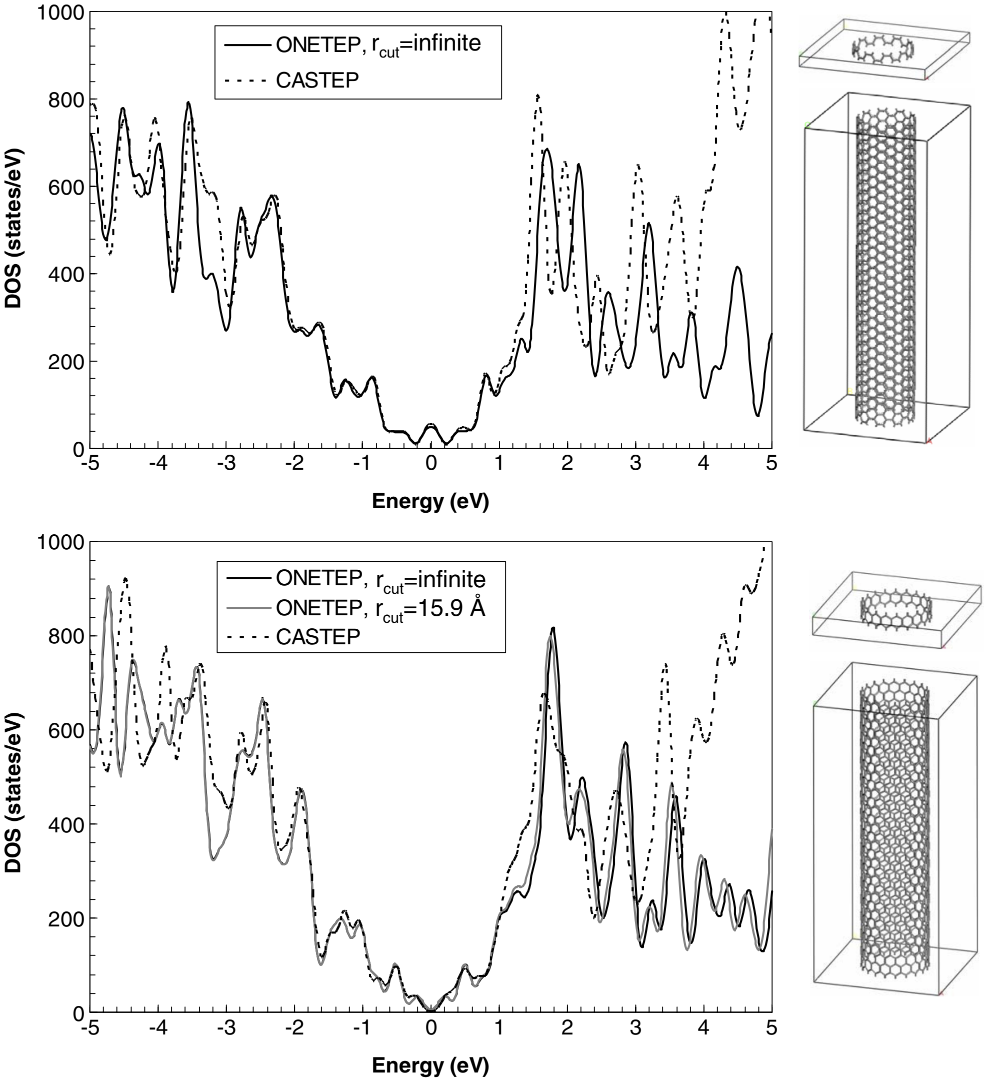

Figure 4. Top panel: the density of states (DOS) of a 30-unit segment of a (10, 10) metallic nanotube as calculated with ONETEP and CASTEP. On the right the ONETEP (30 Å × 30 Å × 73.30 Å) and CASTEP (30 Å × 30 Å × 2.44 Å) simulation cells are shown. Bottom panel: the density of states (DOS) of a 16-unit segment of a (20, 0) semiconducting nanotube as calculated with ONETEP and CASTEP. On the right the ONETEP (30 Å × 30 Å × 67.78 Å) and CASTEP (30 Å × 30 Å × 4.24 Å) simulation cells are shown.

its segments is increased. For example, in figure 4 we show the DOS for the 16 unit segment generated with $r_{\mathrm{cut}} = \infty$ and with $r_{\mathrm{cut}} = 15.9$ Å and the two curves essentially coincide. Our observations are thus consistent with expected behaviour regarding the decay of the density matrix in metallic and non-metallic systems at zero temperature.

ONETEP calculations with $r_{\mathrm{cut}} = \infty$, while not linear scaling, are still perfectly feasible. In particular, most computationally intensive steps such as the construction of the Hamiltonian matrix in the NGWF representation, the construction of the electronic charge density and the calculation of the derivatives of the NGWFs with respect to the expansion coefficients in the psinc basis depend only on the NGWF localization sphere radii r_{loc} and are always perfectly linear scaling independently of the value of r_{cut}. The only step that stops being linear scaling when the density kernel $\mathbf{K}$ is no longer sparse is the optimization of $\mathbf{K}$ which is carried out by using variants of the Li–Nunes–Vanderbilt [54] method and Haynes' [55] penalty functional method which involve matrix multiplications.

It is also worth noting that unlike conventional plane wave approaches where the memory and computation grows with the entire volume of the simulation cell without distinction between vacuum and atomic regions, ONETEP uses algorithms [24, 56] which avoid computation and storage in vacuum regions, thus making possible calculations in very large simulation cells as in this section.

4.2. Solids: crystalline silicon

Here we examine properties of pure crystalline silicon as calculated by ONETEP and CASTEP. For these calculations we have used the LDA with a norm-conserving pseudopotential and plane wave and psinc kinetic energy cut-offs of 283 eV. A cubic unit cell of 1000 atoms was used in the ONETEP calculations and a cubic unit cell of eight atoms was used in the CASTEP calculations, with an equivalent 5 × 5 × 5 **k**-point mesh. The two cells are shown in figure 5. We should note that in a code like CASTEP there are two ways to define the basis set while varying the energy with respect to the lattice parameter. One can either keep the kinetic energy cut-off E_{cut} constant or keep the number of plane wave basis functions N_{PW} constant. The latter approach is conceptually closer to the way the ONETEP calculations are performed in these cases as it is the number of psinc functions that is kept constant, which is equivalent to keeping constant the number of plane waves in the cube of figure 2. Furthermore, in the ONETEP calculations, when varying the lattice parameter, it is important to scale r_{loc} and r_{cut}

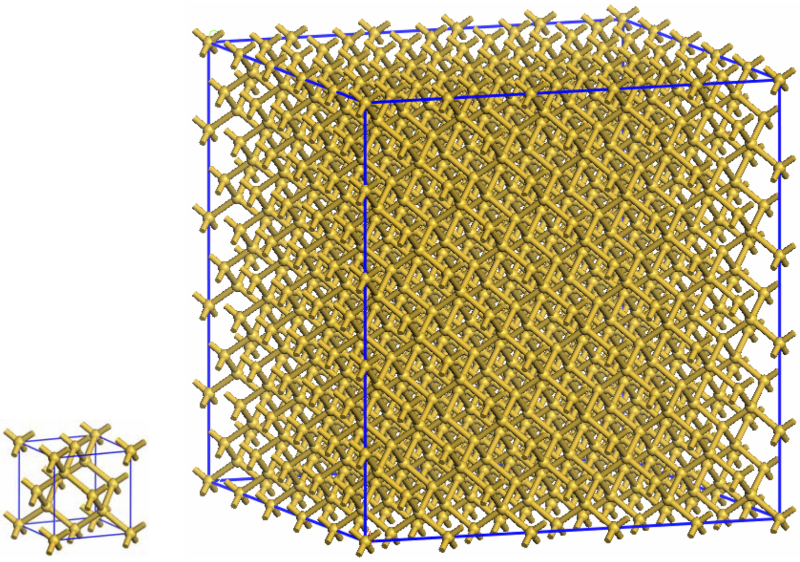

Figure 5. Periodic crystalline silicon. Left: the eight-atom cubic simulation cell used in the calculations with CASTEP. Right: the 1000-atom cubic simulation cell used in the calculations with ONETEP.

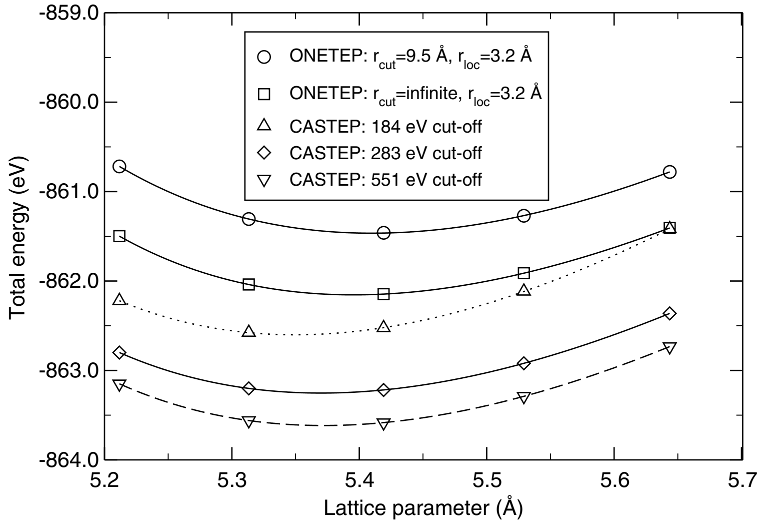

Figure 6. The total energy per eight-atom unit cell of silicon as a function of the lattice parameter for calculations with CASTEP and ONETEP.

proportionately. Throughout this section the values we report for these quantities correspond to a lattice parameter of 5.43 Å.

In figure 6 we show CASTEP constant-N_{PW} plots of the total energy per eight-atom cell as a function of the lattice parameter for kinetic energy cut-offs which correspond to the 'upper bound' (184 eV), 'equivalent' (283 eV) and 'lower bound' (551 eV) cases of figure 2. The two ONETEP curves lie higher in energy than the CASTEP curves because the NGWF radii we used were only 3.2 Å and the total energy is not yet completely converged with respect to them. Nevertheless, the physical properties that we calculate are already converged to a very satisfactory level.

By fitting the calculated energies as a function of the lattice parameter to the Birch–Murnaghan equation of state [57] we obtained values for the lattice constants and bulk moduli of crystalline silicon which we show in table 4. There is excellent agreement between the ONETEP and CASTEP constant-N_{PW} results at 283 eV. For the case of the infinite r_{cut} the lattice constants agree to 0.5% and the bulk moduli to 3.6% while for the case of the 9.5 Å r_{cut} the lattice constants agree to 0.8% and bulk moduli to 3.5%.

The bulk modulus is a quantity which is sensitive to calculation parameters and difficult to converge. Even between the CASTEP calculations with the highest cut-off of 551 eV there remains a difference of 0.7% between the bulk modulus values obtained with constant E_{cut} and constant N_{PW} while the lattice constant difference in this case is reduced to only 0.02%.

4.3. Biomolecules: breast cancer susceptibility proteins

Biomolecules are generally too large for conventional DFT calculations. Nevertheless, a number of insightful studies have been carried out where a small fragment can be isolated from the rest of the biomolecule [58, 59]. Obviously this approach cannot be applied in cases where the interactions extend over a large area, e.g., the case of two large proteins bound to each other. ONETEP can offer great advantages in the study of such molecules since it allows one to perform calculations either on entire biomolecules or at least on segments large enough

Table 4. Lattice constant and bulk modulus of perfect crystalline silicon as calculated by CASTEP, ONETEP and experiment.

Method	Kinetic energy cut-off (eV)	Lattice constant (Å)	Bulk modulus (GPa)
CASTEP, constant E_{cut}	184	5.410	94.7
	283	5.392	94.4
	551	5.383	95.9
CASTEP, constant N_{PW}	184	5.359	109.1
	283	5.380	96.1
	551	5.382	96.6
ONETEP, constant N_{psinc}, $r_{\mathrm{cut}} = \infty$	283	5.406	99.6
ONETEP, constant N_{psinc}, $r_{\mathrm{cut}} = 9.5$ Å	283	5.421	99.5
Experiment		5.430	100.0

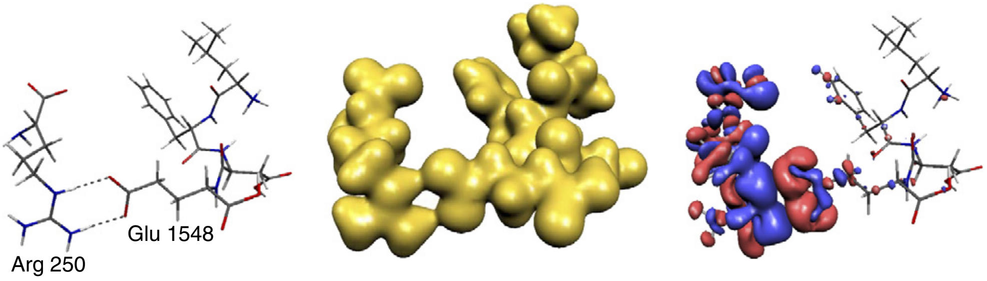

Figure 7. 97-atom segment which includes the bonding interactions between the Arg 250–Glu 1548 residues of the BRCA2-RAD51 complex. Left: stick model of the atomic structure. Middle: isosurface of the electronic change density at a value of 0.02 e^-/a_0^3 from the ONETEP calculation. Right: isosurface of the electronic charge density difference due to bonding at a value of 0.00075 e^-/a_0^3 from the ONETEP calculation.

to contain the entire area of interaction. An example of the latter case is the RAD51–BRCA2 protein complex, for which we present preliminary results in this section.

The breast cancer susceptibility protein [60] BRCA2 regulates the function of RAD51, an enzyme involved in DNA recombination. Crucial to this process is the specific interaction between RAD51 and a BRC motif (sub-region) in BRCA2. There are eight slightly different versions of the BRC motif in a single BRCA2 protein and each of these motifs can interact with a RAD51 protein. Recently the structure of RAD51 bound to one of the BRC motifs (BRC4) has been elucidated by high resolution x-ray diffraction, revealing in a qualitative manner the nature of the interactions at the site of contact between the two proteins [61]. Amino acids with both polar and hydrophobic side chains are involved in these interactions. With this crystal structure as our starting point, we have used calculations with ONETEP to predict the strength of the binding between the two proteins. The 988-atom protein segment we have studied here (figure 8) consists of the entire BRC4 motif and only the A5 α-helix of the RAD51. According to Pellegrini *et al* [61] the major bonding interaction between A5 and BRC4 is a polar interaction involving hydrogen bonding between the side chains of arginine 250 of A5 and glutamic acid 1548 of BRC4. We have first studied just this interaction in isolation by cutting a very small 97-atom segment from the crystal structure of the proteins (figure 7) which contains the relevant amino acids Arg 205 and Glu 1548. The two hydrogen bonds between the cationic side chain of the arginine and the anionic chain of the glutamic acid of

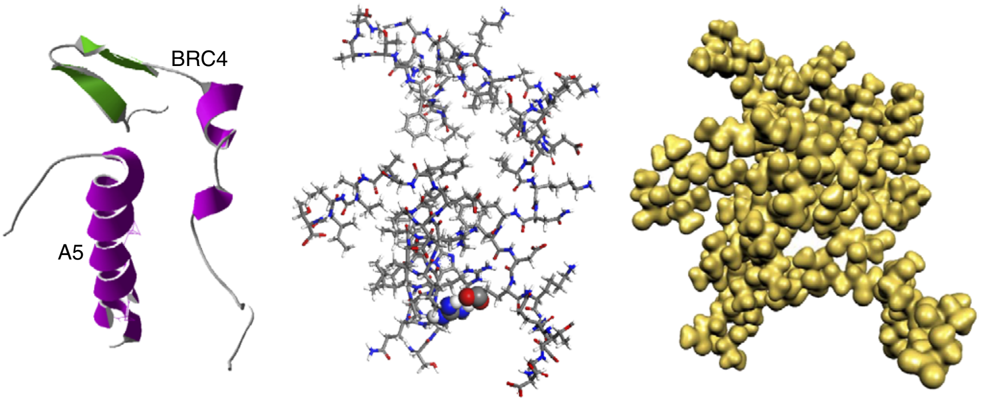

Figure 8. The 988-atom A5-BRC4 complex. Left: tertiary structure. Middle: stick model in atomic detail with the side groups of Arg 250 and Glu 1548 shown in space-filling form. Right: isosurface of the electronic charge density at a value of 0.02 e^{-}/a_0^3 from the ONETEP calculation.

the segment are depicted in figure 7. For this segment we were able to perform calculations with both ONETEP and CASTEP. We have used norm-conserving pseudopotentials, plane wave and psinc kinetic energy cut-offs of 608 eV, cubic simulation cells of 25 Å × 25 Å × 25 Å and the Perdew–Burke–Ernzerhof (PBE) [62] exchange–correlation functional. The radii r_{loc} of the NGWF localization spheres were set to 3.2 Å for the hydrogen atoms and 3.6 Å for all other atoms. The binding energies between these two fragments as calculated by CASTEP and ONETEP are 2.78 and 2.79 eV respectively. Besides the excellent agreement between the two codes it is worth observing that this binding energy is large in comparison to the energy of two regular hydrogen bonds (which individually range from about 100 to 300 meV). It appears that the bulk of the binding strength comes from the electrostatic interaction between the +1 charge of the arginine side group and the −1 charge of the carboxyl of the glutamic acid. Indeed, the classical electrostatic energy of a system of two point charges of +1 and −1 atomic units separated by the same distance as the centres of the side chains of these amino acids is about 3.10 eV, which is rather close to the calculated binding energy. Calculations with the NWChem code with the 6-31+G* Gaussian basis set and the PBE functional produced a binding energy of 2.87 eV which after the counterpoise correction for BSSE became 2.82 eV, in very good agreement with CASTEP and ONETEP given the level of accuracy that can be reached with a Gaussian basis set of this quality (section 3).

For the BRC4-A5 complex of figure 8 the same calculation parameters as for the 97-atom segment were used except for the orthorhombic 60 Å × 50 Å × 60 Å simulation cell. The binding energy between the whole A5 helix and the BRC4 motif that we obtained from our calculations with ONETEP is 5.67 eV. This is about twice as much as the binding energy of the small segment of figure 7 and it shows that the remaining interactions between A5 and BRC4, though small individually, cannot be neglected.

5. Conclusions

In comparison with two well established cubic-scaling density functional methods, we have demonstrated that ONETEP can routinely achieve the highest levels of accuracy that are possible with these methods. Amongst the factors that make this possible is the fact that in ONETEP the calculated properties converge rapidly with the radii of the localization spheres of non-orthogonal generalized Wannier functions (NGWFs) and the rate of self-consistent convergence is affected neither by the size of these regions nor the number of atoms. Next we have demonstrated the wide applicability of the method by presenting exploratory calculations in systems of about 1000 atoms from a wide variety of materials. We have studied semiconducting and metallic nanotubes, crystalline silicon, and the complex of two bound proteins that can play a role in the development of breast cancer. In all these cases we have managed to obtain excellent agreement with CASTEP in comparing either smaller systems of the same material or, where possible by the use of **k**-points, systems of equivalent size. These results confirm that ONETEP is a robust, highly accurate linear-scaling density functional approach, which makes possible a whole new level of large scale simulation in systems of interest to nanotechnology, biophysics and condensed matter physics.

Acknowledgments

C-KS would like to thank the Royal Society for a University Research Fellowship. PDH would like to thank Sidney Sussex College Cambridge for a Research Fellowship. AAM would like to thank Christ's College Cambridge for a Research Fellowship. We are indebted to Professor Ashok Venkitaraman for introducing us to the BRCA2-RAD51 problem. C-KS would like to thank Pamela Schartau for useful discussions on the theory of protein structure. The computing resources of the Cambridge–Cranfield High Performance Computing Facility have been essential for this work.

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] Martin R M 2004 *Electronic Structure. Basic Theory and Practical Methods* (Cambridge: Cambridge University Press)
[4] Artacho E, Machado M, Sánchez-Portal D, Ordejón P and Soler J M 2003 *Mol. Phys.* **101** 1587
[5] Galli G 1996 *Curr. Opin. Solid State Mater. Sci.* **1** 864
[6] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
[7] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Chem. Phys.* **122** 084119
[8] Kohn W 1959 *Phys. Rev.* **115** 809
[9] Des Cloizeaux J 1964 *Phys. Rev.* **135** A685
[10] Baer R and Head-Gordon M 1997 *Phys. Rev. Lett.* **79** 3962
[11] Ismail-Beigi S and Arias T A 1999 *Phys. Rev. Lett.* **82** 2127
[12] He L and Vanderbilt D 2001 *Phys. Rev. Lett.* **86** 5341
[13] Skylaris C-K, Mostofi A A, Haynes P D, Diéguez O and Payne M C 2002 *Phys. Rev.* B **66** 035119
[14] McWeeny R 1960 *Rev. Mod. Phys.* **32** 335
[15] Hernández E and Gillan M J 1995 *Phys. Rev.* B **51** 10157
[16] Goringe C M, Hernández E, Gillan M J and Bush I J 1997 *Comput. Phys. Commun.* **102** 1
[17] Haynes P D and Payne M C 1997 *Comput. Phys. Commun.* **102** 17
[18] Fattebert J-L and Bernholc J 2000 *Phys. Rev.* B **62** 1713
[19] Fattebert J-L and Gygi F 2004 *Comput. Phys. Commun.* **162** 24
[20] Pask J E and Sterne P A 2005 *Modelling Simul. Mater. Sci. Eng.* **13** R71
[21] Mostofi A A, Haynes P D, Skylaris C-K and Payne M C 2003 *J. Chem. Phys.* **119** 8842
[22] Baye D and Heenen P-H 1986 *J. Phys. A: Math. Gen.* **19** 2041
[23] Varga K, Zhang Z and Pantelides S T 2004 *Phys. Rev. Lett.* **93** 176403
[24] Mostofi A A, Skylaris C-K, Haynes P D and Payne M C 2002 *Comput. Phys. Commun.* **147** 788
[25] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[26] Sankey O F and Niklewski D J 1989 *Phys. Rev.* B **40** 3979
[27] Kenny S D, Horsfield A P and Fujitani H 2000 *Phys. Rev.* B **62** 4899
[28] Junquera J, Paz O, Sánchez-Portal D and Artacho E 2001 *Phys. Rev.* B **64** 235111

[29] Anglada E, Soler J M, Junquera J and Artacho E 2002 *Phys. Rev.* B **66** 205101

[30] Davidson E R and Feller D 1986 *Chem. Rev.* **86** 681

[31] Bardo R D and Ruedenberg K 1974 *J. Chem. Phys.* **60** 918

[32] Dunning T H Jr 1989 *J. Chem. Phys.* **90** 1007

[33] Segall M D, Lindan P J D, Probert M J, Pickard C J, Hasnip P J, Clark S J and Payne M C 2002 *J. Phys.: Condens. Matter* **14** 2717

[34] Branden C and Tooze J 1998 *Introduction to Protein Structure* 2nd edn (New York: Garland Publishing)

[35] Ceperley D M and Alder B J 1980 *Phys. Rev. Lett.* **45** 566

[36] Perdew J P and Zunger A 1981 *Phys. Rev.* B **23** 5048

[37] Hamann D R, Schlüter M and Chiang C 1979 *Phys. Rev. Lett.* **43** 1494

[38] Lee M-H, Lin J-S, Payne M C, Heine V, Milman V and Crampin S 2004 Kinetic energy tuning for optimising pseudopotenials and projectior reduction *Psi-k Network Scientific Highlight of the Month 12* http://psi-k.dl.ac.uk/index.html?highlights

[39] Milman V and Lee M-H 1996 *J. Phys. Chem.* **100** 6093

[40] Skylaris C-K, Diéguez O, Haynes P D and Payne M C 2002 *Phys. Rev.* B **66** 073103

[41] Bowler D R and Gillan M J 1998 *Comput. Phys. Commun.* **112** 103

[42] Straatsma T P *et al* 2003 *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.5* (Richland, WA: Pacific Northwest National Laboratory)

[43] Hehre W J, Stewart R F and Pople J A 1969 *J. Chem. Phys.* **51** 2657

[44] Binkley J S, Pople J A and Hehre W J 1980 *J. Am. Chem. Soc.* **102** 939

[45] Hehre W J, Ditchfield R and Pople J A 1972 *J. Chem. Phys.* **56** 2257

[46] Clark T, Chandrasekhar J and Schleyer P V R 1983 *J. Comput. Chem.* **4** 294

[47] Hariharan P C and Pople J A 1973 *Theor. Chim. Acta* **28** 213

[48] Dunning T H 1989 *J. Chem. Phys.* **90** 1007

[49] Kendall R A, Dunning T H Jr and Harrison R J 1992 *J. Chem. Phys.* **96** 6769

[50] Vosko S J, Wilk L and Nusair M 1980 *Can. J. Phys.* **58** 1200

[51] Boys S F and Bernardi F 1970 *Mol. Phys.* **19** 553

[52] Saito R, Dresselhaus G and Dresselhaus M S 1998 *Physical Properties of Carbon Nanotubes* (London: Imperial College Press)

[53] Odom T W, Huang J-L, Kim P and Lieber C M 2000 *J. Phys. Chem.* B **104** 2794

[54] Li X P, Nunes R W and Vanderbilt D 1993 *Phys. Rev.* B **47** 10891

[55] Haynes P D and Payne M C 1999 *Phys. Rev.* B **59** 12173

[56] Skylaris C-K, Mostofi A A, Haynes P D, Pickard C J and Payne M C 2001 *Comput. Phys. Commun.* **140** 315

[57] Murnaghan F D 1944 *Proc. Natl Acad. Sci.* **30** 244

[58] Molteni C, Frank I and Parrinello M 1999 *J. Am. Chem. Soc.* **121** 12177

[59] Segall M D 2002 *J. Phys.: Condens. Matter* **14** 2957

[60] Venkitaraman A R 2002 *Cell* **108** 171

[61] Pellegrini L, Yu D S, Lo T, Anand S, Lee M, Blundell T L and Venkitaraman A 2002 *Nature* **420** 287

[62] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865