THE JOURNAL OF CHEMICAL PHYSICS **122**, 084119 (2005)

Introducing ONETEP: Linear-scaling density functional simulations on parallel computers

Chris-Kriton Skylaris,[a] Peter D. Haynes, Arash A. Mostofi, and Mike C. Payne
Theory of Condensed Matter, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 29 September 2004; accepted 4 November 2004; published online 23 February 2005)

We present ONETEP (order-N electronic total energy package), a density functional program for parallel computers whose computational cost scales linearly with the number of atoms and the number of processors. ONETEP is based on our reformulation of the plane wave pseudopotential method which exploits the electronic localization that is inherent in systems with a nonvanishing band gap. We summarize the theoretical developments that enable the direct optimization of strictly localized quantities expressed in terms of a delocalized plane wave basis. These same localized quantities lead us to a physical way of dividing the computational effort among many processors to allow calculations to be performed efficiently on parallel supercomputers. We show with examples that ONETEP achieves excellent speedups with increasing numbers of processors and confirm that the time taken by ONETEP as a function of increasing number of atoms for a given number of processors is indeed linear. What distinguishes our approach is that the localization is achieved in a controlled and mathematically consistent manner so that ONETEP obtains the same accuracy as conventional cubic-scaling plane wave approaches and offers fast and stable convergence. We expect that calculations with ONETEP have the potential to provide quantitative theoretical predictions for problems involving thousands of atoms such as those often encountered in nanoscience and biophysics. © *2005 American Institute of Physics.* [DOI: 10.1063/1.1839852]

I. INTRODUCTION

The equations of quantum mechanics govern the correlated motions of electrons and nuclei and are thus essential in any theoretical description of the chemical or physical properties of matter. Apart from trivial cases, these equations are impossible to solve with pen and paper and highly sophisticated computational methods for their solution have been devised.[1,2] Amongst them the Kohn–Sham density functional theory (DFT) formalism[3,4] for electronic structure calculations has become established as an approach that provides a very good description of electron correlation effects while keeping the size of calculations tractable. DFT calculations have become an indispensable tool for the study of matter with myriads of applications in areas such as chemistry,[5] biochemistry,[6] polymers,[7] and materials[8,9] to name a few. However even DFT calculations suffer from an unfavorable scaling: the time taken to perform such a calculation on a computer increases asymptotically with the cube of the number of atoms. This cubic scaling is a consequence of the delocalized nature of the wave functions which are the eigensolutions of the Kohn–Sham single particle Hamiltonian,[4,10] and limits the number of atoms we can treat to a few hundred at most. There are many exciting problems at the interface between the microscopic and mesoscopic worlds, particularly in the emerging fields of biophysics and nanoscience, whose theoretical investigation would be possible only with an accurate quantum mechanical description of the interactions between thousands of atoms.

In an attempt to extend the application of DFT to such problems, researchers in recent years have put substantial effort into the construction of DFT methods which are *linear-scaling*,[11,12] i.e., with a cost which increases asymptotically only linearly with the number of atoms. These methods exploit the electronic localization[13,14] that is inherent in systems with a band gap and seek to optimize quantities that (in principle) are infinite in extent, but decay exponentially, such as the single-particle density matrix[10] or Wannier functions.[15,16] A common point between these methods is that the onset of linear-scaling occurs only after the number of atoms exceeds a critical value. An important performance characteristic then is the *crossover point*, the number of atoms at which a linear-scaling approach becomes faster than a conventional cubic-scaling approach. This crossover point is system dependent but often lies in the order of hundreds of atoms. As single processor workstations are capable of calculations with roughly no more than 100 atoms, it is important to use multiprocessor (parallel) computers if we are to reap any benefits from linear-scaling DFT. Conversely, we could argue that only linear-scaling methods are suited to take best advantage of parallel computers since, only with them does an eightfold increase in computational power allow calculations for eight times as many atoms instead of only twice as many atoms as in conventional approaches. It is not surprising therefore that the development of linear-scaling methods has often advanced hand in hand with the

[a] Author to whom correspondence should be addressed. Present address: Department of Physical and Theoretical Chemistry, South Parks Road, Oxford OX1 3QZ, UK.
Electronic mail: chris-kriton.skylaris@chem.ox.ac.uk
URL: http://www.chem.ox.ac.uk/researchguide/ckskylaris.html

development of suitable algorithms for calculations on parallel computers.[17–20]

To be useful, a linear-scaling method should have a systematic way to reduce error to any value desired, in the same way as conventional methods. However progress towards this goal has been slow as it has been difficult to devise generally applicable schemes to truncate the exponentially decreasing "tails" of the density matrix or Wannier functions while maintaining control over the accuracy or the stability of the iterative minimization procedure. Most linear-scaling approaches use *nonorthogonal localized* basis sets to express their (also localized) functions. These approaches can be classified into methods which use atomic-like basis sets such as Gaussian functions,[21] Slater functions,[22] spherical Bessel functions[23] or numerical atomic orbitals,[24] and methods which use simpler localized basis sets such as polynomials[25] or real-space grids.[19,26]

Our linear-scaling method[27] is different from all other approaches as it uses a basis set of highly localized functions which are *orthogonal*. This approach allows for systematic control of truncation errors and is compatible with an accurate representation of the kinetic energy operator,[28] which ensures variational behavior with respect to the basis set.[29] Our linear-scaling method is implemented in ONETEP (order-*N* electronic total energy package) which has been developed with algorithms intended for calculations on parallel supercomputers and is the subject of this paper.

We give a brief presentation of the formalism of linear-scaling methods in Sec. II. In Sec. III we focus on ONETEP, and explain its capabilities with theoretical arguments and example calculations. In Sec. IV we give an overview of the principles behind the parallel implementation of ONETEP which is based again on the real space localization. Finally in Sec. V we demonstrate how ONETEP takes advantage of parallel computers in order to perform calculations with thousands of atoms.

II. THEORETICAL BACKGROUND

Our aim is to solve a set of single-particle Schrödinger equations in a potential $V(\mathbf{r})$, as is the case in DFT

$$\hat{H}\psi_i(\mathbf{r})=\left[-\frac{\hbar^2}{2m}\nabla^2+V(\mathbf{r})\right]\psi_i(\mathbf{r})=\epsilon_i\psi_i(\mathbf{r}), \tag{1}$$

where $\hat{H}$ is the single-particle Hamiltonian of the system with energy eigenvalues ϵ_i and corresponding spatial eigenfunctions (also known as "orbitals") $\psi_i(\mathbf{r})$ which are orthogonal.

All the information about the ground state of our system is contained in the *single-particle* density matrix $\rho(\mathbf{r},\mathbf{r}')$,

$$\rho(\mathbf{r},\mathbf{r}')=\sum_i f_i\psi_i(\mathbf{r})\psi_i^*(\mathbf{r}'), \tag{2}$$

where f_i is the occupancy of state $\psi_i(\mathbf{r})$ and at zero temperature it is restricted to either 0 or 1. The charge density $n(\mathbf{r})$, which is the central quantity in DFT, is given by the diagonal elements of the density matrix

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

where the factor of 2 above is included to account for electronic spin as we assume here a closed shell description. Provided there is a band gap in the system, the density matrix (2) decays exponentially[30–32] as a function of the distance between $\mathbf{r}'$ and $\mathbf{r}$. This property can be exploited to truncate the density matrix to a sparse band-diagonal form such that the amount of information it contains increases linearly with its size. To achieve this in practice, the density matrix is expressed in the equivalent form (throughout this paper a summation will be implied over repeated Greek indices)

$$\rho(\mathbf{r},\mathbf{r}')=\phi_\alpha(\mathbf{r})K^{\alpha\beta}\phi_\beta^*(\mathbf{r}'), \tag{4}$$

where the $\{\phi_\alpha\}$ are a set of *spatially localized, nonorthogonal* basis functions and the matrix $\mathbf{K}$, as defined by the above equation, is called the *density kernel*.[33] This form allows for a practical, "coarse-grained" truncation of the density matrix through truncation of the density kernel. Thus we ensure the density kernel is sparse by enforcing the condition

$$K^{\alpha\beta}=0 \quad \text{when } |\mathbf{R}_\alpha-\mathbf{R}_\beta|>r_{\text{cut}}, \tag{5}$$

where $\mathbf{R}_\alpha$ and $\mathbf{R}_\beta$ are the "centers" of the localization regions of the functions $\phi_\alpha(\mathbf{r})$ and $\phi_\beta(\mathbf{r})$.

Often a linear combination of atomic orbitals (LCAO) approach is followed where the basis $\{\phi_\alpha\}$ consists of atomic orbitals. Their radial shapes can be expanded in spherical Bessel functions,[23] Gaussians[34,35]—where sparsity is commonly imposed via "thresholding" [36] rather than by Eq. (5)—and numerical atomic orbitals as in the SIESTA program[37] where instead of the density kernel, orthogonal Wannier-like functions are truncated. All these sets of functions are taken preoptimized and remain fixed during the calculation. Based on only operations with a linear cost such as the construction of the Hamiltonian matrix in the LCAO basis and sparse matrix algebra, a number of efficient techniques[38–41] have been developed that minimize the energy while satisfying the difficult nonlinear constraints of density matrix idempotency or Wannier-like function orthogonality.[42]

The main concern with approaches of the LCAO type is the *transferability* of the basis set. Even with the available recipes for the generation of high quality atomic orbitals,[24,43–45] the number of such functions per atom can be large, and a good level of expertize is needed to generate a basis set of the size and type that balances efficiency and required accuracy for each new problem. The size of sparse matrices for a given r_{cut} increases with the square of the number of atomic orbitals per atom while the operation cost (prefactor) for linear-scaling matrix multiplications increases with the cube. As a rule, preliminary calculations with a number of basis sets are performed to select the most suitable one and "calibrate" the method. This is in contrast to the "black box" behavior of the plane wave approach where systematic improvement of the basis is guaranteed by increasing a single parameter. Hence, while low level LCAO calculations are relatively easy to do, improving the accuracy quickly becomes both technically demanding and computationally very expensive.

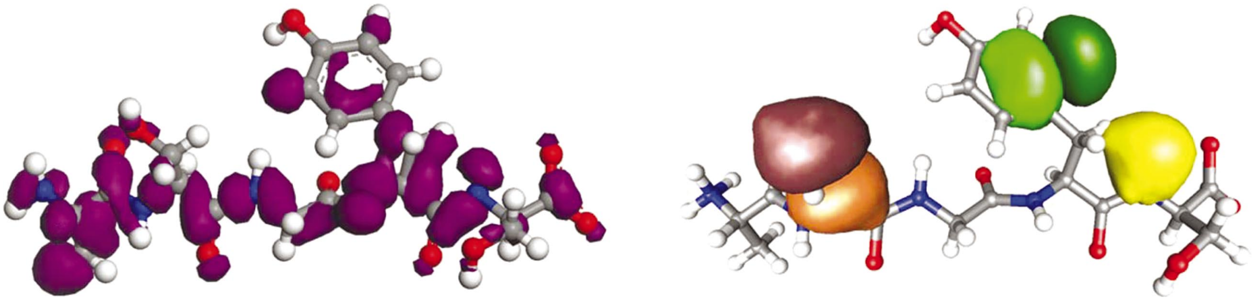

FIG. 1. (Color) Left, one delocalized orbital $\psi_i(\mathbf{r})$ from a conventional DFT calculation with the CASTEP code (Ref. 50) on a peptide. Right, three optimized NGWFs $\phi_\alpha(\mathbf{r})$, $\phi_\beta(\mathbf{r})$, and $\phi_\gamma(\mathbf{r})$ from a ONETEP calculation on the same peptide.

III. OVERVIEW OF ONETEP

The state of affairs in ONETEP is different from other linear-scaling approaches. We overcome the matrix size problem by using a minimal number of $\{\phi_\alpha\}$ localized functions per atom and address the transferability issue by optimizing these functions *during* the calculation. Therefore the $\{\phi_\alpha\}$ are no longer our (atomic orbital) basis set, rather they are quantities to be determined during the calculation along with the density kernel **K**. We call the $\{\phi_\alpha\}$ nonorthogonal generalized Wannier functions (NGWFs) (Ref. 27) (see Fig. 1). We enforce strict localization on our NGWFs by confining them to spherical regions centered on atoms and by constantly truncating any contributions that may develop outside these *localization spheres* during our *conjugate gradients* optimization[46] procedure. To achieve this, we expand them in a basis of periodic sinc[47] or psinc[48] functions $\{D_k(\mathbf{r})\}$:

$$\phi_\alpha(\mathbf{r}) = \sum_k D_k(\mathbf{r}) C_{k,\alpha}. \tag{6}$$

Each psinc is a highly localized spike-like function and the index k indicates the grid point on which $D_k(\mathbf{r})$ is centered as the set of psincs covers a regular grid of points throughout the volume of the simulation cell. Each $\phi_\alpha(\mathbf{r})$ then is confined in its localization sphere of radius r_{loc}^α centered on an atom, by truncating its expansion in Eq. (6), as shown in Fig. 2. In general, for an arbitrary localized basis set, such an act of truncation would lead to a breakdown of the conjugate gradients minimization schemes employed in electronic structure calculations.[46] Only for the case of an orthogonal basis set are the gradient contributions inside and outside the localization sphere decoupled[49] so that the *selective optimization* of quantities inside the localization sphere is stable. By construction, the psinc basis set is orthogonal.

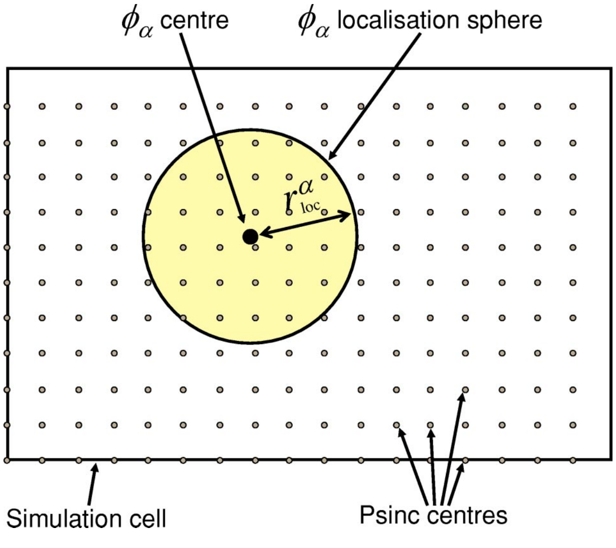

FIG. 2. Imposing localization on the $\phi_\alpha(\mathbf{r})$ function in real space. From the regular grid of psinc functions $D_k(\mathbf{r})$, only the ones within its localization sphere are allowed to contribute to $\phi_\alpha(\mathbf{r})$.

The psinc functions, through which all quantities are ultimately expressed in ONETEP, are connected to plane waves by means of a Fourier transform. Due to this property ONETEP is essentially a linear-scaling reformulation of the plane wave pseudopotential DFT approach. The quality of the psinc basis set can be systematically improved by varying only one parameter, the grid spacing of the psincs, which is equivalent to the kinetic energy cutoff of the plane waves. The equivalence of our method with the conventional plane-wave pseudopotential approach can be best demonstrated by example. We have chosen here the case of the hydrogen bond formed by two water molecules as a rather challenging test involving a weak chemical bond, close to the limits of the accuracy of DFT. In Fig. 3 we plot the energy as a function of the bond distance. Calculations with ONETEP and with the conventional plane wave pseudopotential approach as implemented in the CASTEP code[50] were carried out using the same norm-conserving pseudopotentials and plane waves up to the

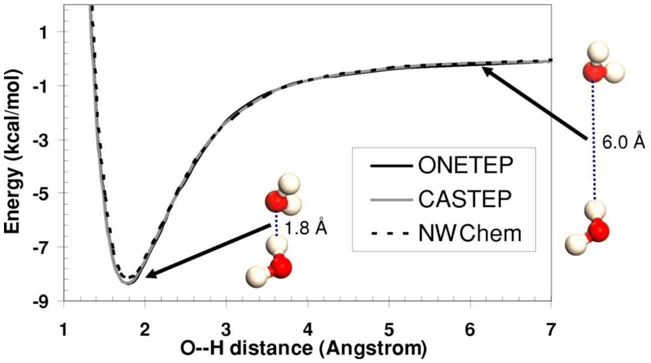

FIG. 3. The potential energy curve of two hydrogen bonded water molecules as a function of H-bond distance calculated with ONETEP (this work), CASTEP (conventional plane wave pseudopotential approach), and NWChem (Gaussian basis all-electron approach).

same kinetic energy cutoff of 95 Ry. The NGWF localization sphere radii $r_{\mathrm{loc}}^{\alpha}$ in ONETEP were set to 3.3 Å. There is excellent agreement between ONETEP and CASTEP as the two curves essentially coincide. The equilibrium bond length and the curvature of the curve at this length as determined by ONETEP differ from the CASTEP results by only 0.3% and 0.5%, respectively. In the same figure we also show the potential energy curve obtained with the all-electron Gaussian basis function code NWChem (Ref. 51) using the cc-pVTZ basis[52] augmented with its corresponding set of diffuse functions.[53] This substantial basis set is necessary here to describe accurately the weak interactions due to the hydrogen bond and comprises 280 contracted Gaussian atomic orbitals. In contrast, the ONETEP calculation uses only 12 NGWFs (four on each oxygen atom and one on each hydrogen atom)—these numbers show how large the difference in matrix sizes in accurate calculations with ONETEP and LCAO-type codes can be. Given the fact that NWChem performs all-electron calculations, the agreement with ONETEP is extremely good: the equilibrium bond length and the curvature of the curve at this length as determined by ONETEP differ from the NWChem results by 0.6% and 2.3%, respectively. It is also worth observing from Fig. 3 the smoothness of the ONETEP curve. This is a consequence of the strict mathematical consistency of all operations in ONETEP, such as the fact that because of the plane wave nature of our basis we are able to calculate both the kinetic and Hartree energies using the same Fourier transform methods.[54] This, combined with the fact that the psinc functions are fixed in space which means that they do not involve so-called ‘‘Pulay forces,’’ [55] greatly facilitates the essential calculation of forces on atoms irrespective of their position.

A known difficulty of self-consistent calculations is that the number of iterations needed to reach a given convergence threshold per atom can be very large, and can often be exacerbated by large basis sets. Even in methods such as ONETEP where the computational cost of each NGWF conjugate gradients iteration is linear-scaling, the number of such iterations can be so large that self-consistent minimization is prohibitively inefficient. To overcome this obstacle, we have developed a preconditioning scheme[48] which enables our calculations to converge in a small number of iterations (typically 20–40) which is independent of the number of atoms. We will return to this point in Sec. V.

IV. PARALLELIZATION STRATEGY

Our formulation of the ONETEP algorithm is similar to that presented in our earlier work[27] with a few exceptions noted below. We shall only review the parts relevant to the implementation on parallel computers here; for full details we refer the reader to our earlier paper.[27] Furthermore, here we seek to give the reader a general overview of concepts rather than an exhaustive description of algorithms which we leave for another, more technical paper.

We use the Message Passing Interface (MPI) library for communication between processors[56,57] and note that in this parallelization approach each processor possesses its own independent portion of the data. Our parallelization strategy requires distribution across processors both of the computational effort and the data. Our model is such that all processors perform equivalent tasks at all times and as a result our parallel code can also run on computers with a single processor without any modification. In our discussion we will use N_P to represent the total number of processors, numbered from 0 to $(N_P - 1)$.

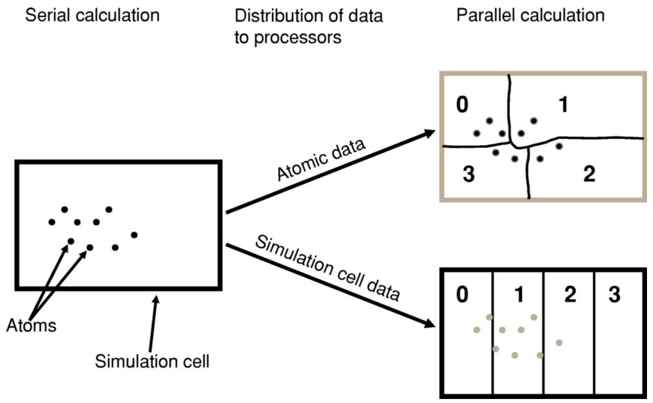

FIG. 4. Schematic two-dimensional example of our data parallelization strategy. For clarity only four processors are shown (numbered from 0 to 3) and nine atoms which are represented as dots. The distribution of atomic data to processors involves partitioning the material into fragments made up of atoms in close proximity. The distribution of simulation cell data involves partitioning the simulation cell into slabs.

A. Data parallelization

Two types of data are parallelized in ONETEP. First, there is data directly associated with each atom such as the expansion coefficients $C_{k,\alpha}$ of the NGWFs in the psinc basis according to Eq. (6). The number of such coefficients increases linearly with the number of atoms, and since our basis set consists of a large number of highly localized psinc functions distribution of such quantities is essential. Secondly, there is data whose size is proportional to the volume of the simulation cell, such as the charge density and the local potential. While their size formally does not depend on the number of atoms, in practice larger simulation cells are needed to accommodate increasing numbers of atoms and soon distribution of simulation cell related quantities becomes essential. Figure 4 illustrates our parallelization strategy for *atomic data* and for *simulation cell data*.

Our parallelization strategy for the atomic data takes advantage of the strict localization of the NGWFs. Each processor P is assigned a number of atoms $N_{\mathrm{at}}^{(P)}$ which is a subset of the total number of atoms N_{at} in the system. The distribution of atoms is performed so that the number of NGWFs $N_{\mathrm{NGWF}}^{(P)}$ on each processor is approximately the same in order to achieve balance in the load of the computation. Another important issue is the minimization of the necessary communication between processors. As shown in Fig. 4, we desire the atoms allocated to a processor to be in close proximity so that the number of their NGWF localization sphere overlaps with those of atoms belonging to other processors is as small as possible. This, in turn, minimizes the number of NGWFs that need to be communicated from one processor to another when computing quantities such as the Hamiltonian matrix in the NGWF representation $H_{\alpha\beta}=\langle\phi_\alpha|\hat{H}|\phi_\beta\rangle$. To achieve this goal we create a Peano "space filling" fractal curve based on which we rearrange the atoms according to their proximity in space.[18] A further positive outcome from this technique is that it leads to clustering of nonzero values near the diagonal of our sparse matrices.

The distribution of simulation cell related quantities such as the charge density to processors is more straightforward, as shown in Fig. 4. The simulation cell is partitioned into slabs along one of its dimensions and each processor is allocated the quantities that belong to its slab.

B. Parallelization of operations and communication

The distribution of data described in the preceding section allows the division of the computational work among the processors. The bulk of the computation goes first into the calculation of the total electronic energy

$$E[\{K^{\alpha\beta}\},\{\phi_\alpha\}]=2K^{\alpha\beta}H_{\beta\alpha}+E_{\mathrm{DC}}[n], \tag{7}$$

where the first term is the band structure energy and the second term is the "double-counting"[11] correction which contains the exchange-correlation energy and terms compensating for spurious interactions contained in the first term. Second, but just as demanding from a computational viewpoint, we have the calculation of the gradient of the energy with respect to the NGWFs in the psinc basis

$$\frac{\delta E}{\delta\phi_\alpha(\mathbf{r})}=4\,\hat{H}\phi_\beta(\mathbf{r})K^{\beta\alpha}. \tag{8}$$

ONETEP is designed so that the number of operations to calculate these quantities increases asymptotically only linearly with the number of atoms.[27] As atomic data are distributed, communication between processors is required, and the energy of Eq. (7) is calculated as a sum of contributions from each processor

$$\begin{aligned}E[\{K^{\alpha\beta}\},\{\phi_\alpha\}]&=\sum_{P=0}^{N_P-1}E^{(P)}[\{K^{\alpha\beta}\},\{\phi_\alpha\}]\\&=\sum_{P=0}^{N_P-1}\left[\sum_{P'=0}^{N_P-1}K^{\alpha\beta}\langle\phi_\beta^{(P')}|\hat{H}|\phi_\alpha^{(P)}\rangle\right.\\&\qquad\left.+E_{\mathrm{DC}}[n^{(P)}]\right],\end{aligned} \tag{9}$$

where each function $\phi_\beta^{(P')}(\mathbf{r})$ from a processor $P'\neq P$ must be communicated to P, provided its localization sphere overlaps with the sphere of $\phi_\alpha^{(P)}(\mathbf{r})$. The second term of Eq. (9) is calculated entirely on P from its slab of the charge density $n^{(P)}$, with the exception of the exchange-correlation energy for the case of generalized gradient approximation (GGA) functionals where some communication between the processors is required to obtain the gradient of the charge density $(\boldsymbol{\nabla}n)^{(P)}$. A related approach is followed in the calculation of the NWGF gradient in Eq. (8) where each processor only computes and stores the gradient relevant to its functions $\delta E/\delta\phi_\alpha^{(P)}(\mathbf{r})$.

A suitable communication model for the above tasks should allow for pairwise "point-to-point" communication between all distinct pairs of processors. A further demand is that the model must allow processors uninterrupted computation while sending and receiving only the NGWFs needed due to overlapping localization spheres. We have developed an efficient communication algorithm which is scalable in principle to an arbitrary number of processors. Our communication model consists of $2N_P-1$ steps and is outlined in Fig. 5. Again, for the sake of conciseness, we focus here on one specific example with four processors, therefore we have $2\times4-1=7$ steps. At each step we show four boxes numbered from 0 to 3, representing the four processors. The arrows connecting them indicate the direction of point-to-point communication. Next to each communication step we show a 4×4 matrix whose elements represent all the possible pairs of processors. The shaded elements represent the computation taking place at each step: the column is the processor that performs the computation while receiving data from the processor of the corresponding row. Step 1 always involves the diagonal elements of the matrix and thus no communication. As a consequence of our parallelization strategy for atomic data, the matrix of Fig. 5 has increasing sparsity away from the diagonal. Our algorithm takes this feature into account and communicates data only when it is required for computation.

Our parallelization strategies for atomic data and simulation cell data cannot be independent of each other; for example, the calculation of the Hartree (Coulomb) potential contribution to the $H_{\alpha\beta}$ matrix requires operations between atomic data such as the $\{\phi_\alpha\}$ and simulation cell data such as the Hartree potential $\hat{V}_H(\mathbf{r})$. These operations are performed in subregions of the simulation cell which are independent of system size by means of the FFT-box technique which allows us to retain an accurate representation of quantum mechanical operators and their properties.[28,54]

V. LINEAR-SCALING WITH PROCESSORS AND WITH ATOMS

ONETEP is a general purpose electronic structure code and as such it should be able to take advantage of parallel computers in all potential applications. While ONETEP has all the familiar characteristics of the plane wave approach (systematic basis set, periodic boundary conditions, pseudopotentials) most of the computation is done in real space with localized functions. Based on these considerations, the parallel algorithms we have described in Sec. IV are intended to be scalable with the size of the calculation to an arbitrary number of processors. In practice we need to have more than one atom per processor for the communication not to dominate the total computational time. However, we have observed that only ten atoms per processor $N_{\mathrm{at}}^{(P)}$ is already enough for good parallel scaling in most cases. All the calculations we report here were performed on the Sun Fire 15K parallel supercomputer of the Cambridge-Cranfield high performance computing facility (CCHPCF).

A straightforward way to assess the performance of our code on parallel computers is by measuring the speedup of

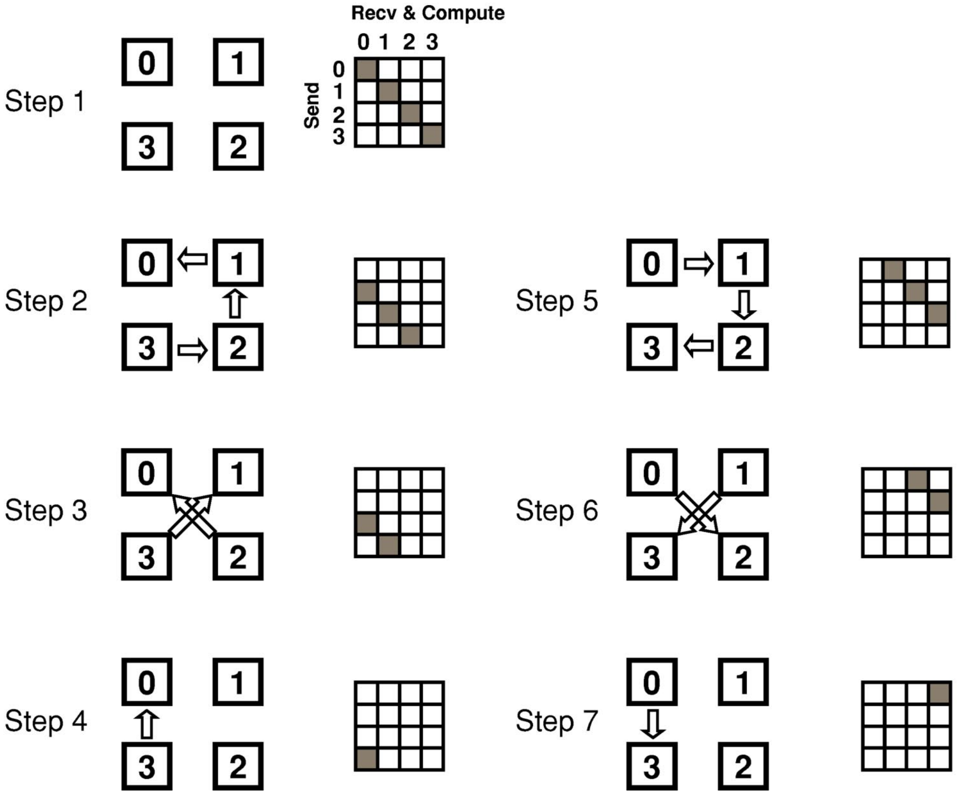

FIG. 5. Our point-to-point communication model. Only four processors are shown for clarity (numbered from 0 to 3). The arrows indicate the direction of communication at each step. The column of each shaded element shows the processor performing computations while receiving data from the processor of the corresponding row.

the computational time on increasing the number of processors. In Fig. 6 we show the speedups we obtain for calculations run on 4, 8, 16, 32, 48, and 64 processors. We focus here on two examples, an 800-atom chiral boron nitride nanotube and a 1403-atom polyglycine peptide in a globular conformation. We observe that both curves show an almost linear speedup up to 64 processors. The speedups we achieve remain substantial even when we get to 64 processors with 79% of the ideal value for the case of the nanotube and 72% in the case of polyglycine. The very regular structure of the nanotube leads to an ideal partitioning of the atomic data by our parallelization strategy to achieve a near-optimal balance of the computation and communication load, hence the parallel speedup in this case is greatest. The irregular three-dimensional structure of the polyglycine is a challenging test for our parallel algorithms. While these irregularities are reflected in the jumps that the polyglycine curve in Fig. 6 displays with the increasing number of processors, it is particularly pleasing to note that the speedups remain high through its range, from which we can conclude that the distribution of atoms to processors and hence the balancing of tasks is still done in a satisfactory way.

In addition to the linear decrease of the time for a given calculation as a function of increasing the number of processors, the other significant performance advantage of a code such as ONETEP is the linear-scaling of the total time with the

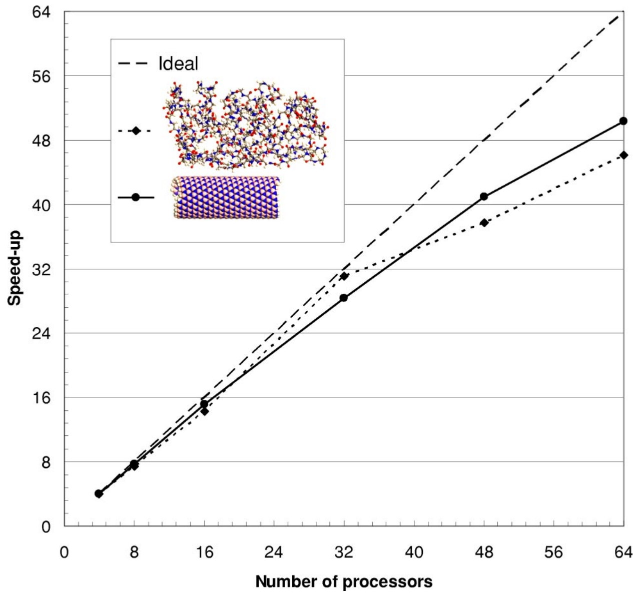

FIG. 6. Parallel scalability tests of ONETEP on the SUN Fire 15K supercomputer of the CCHPCF. The speedup for the time taken for a single NGWF iteration is plotted as a function of the number of processors for a polyglycine molecule (broken line) and a boron nitride nanotube (solid line).

TABLE I. Total energy calculations with ONETEP on pieces of DNA with 64 processors. The time taken in hours is shown as a function of the number of atoms, and equivalently, base pairs. Also shown are the number of NGWF iterations needed to converge and the final convergence of the energy per atom.

Atoms	Base pairs	Total time (h)	Iterations	$\Delta E/\text{atom}$ (E_h)
117	2	2.0	26	1.3×10^{-8}
379	6	6.4	29	6.6×10^{-9}
641	10	11.2	29	7.6×10^{-9}
1296	20	26.1	30	1.1×10^{-8}
1951	30	42.8	29	1.1×10^{-8}
2606	40	56.9	31	6.6×10^{-9}

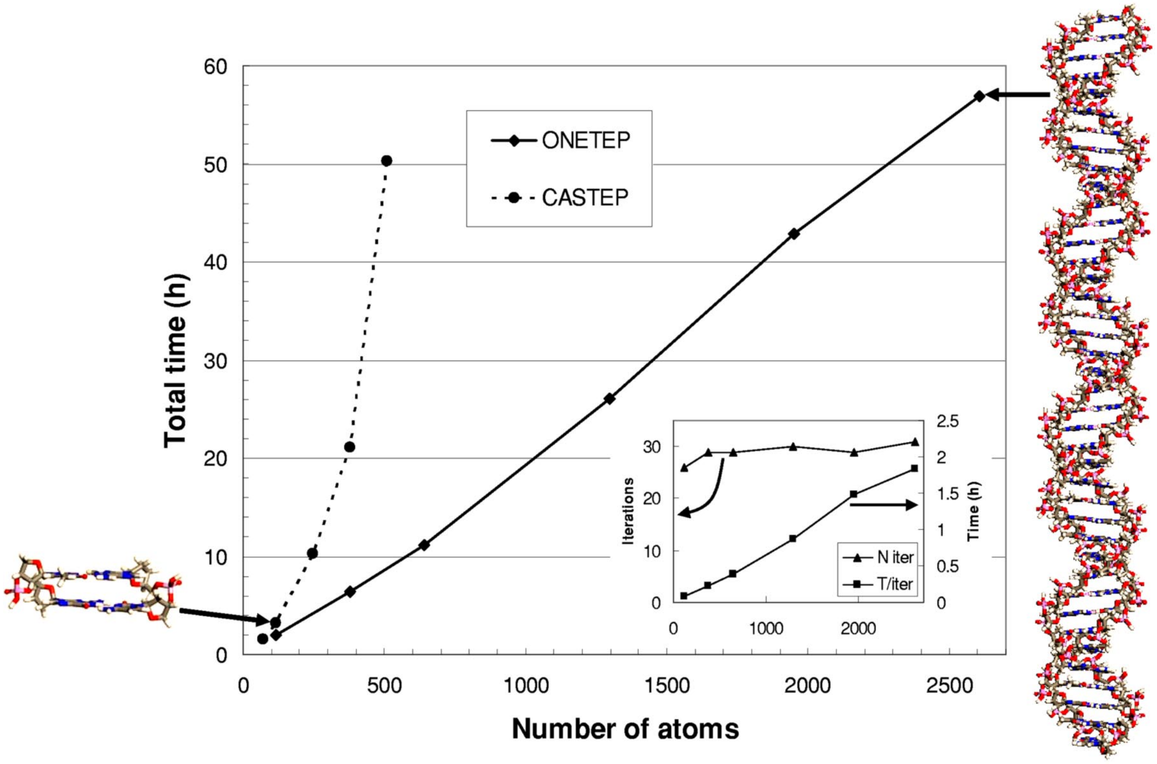

FIG. 7. Total energy calculations with ONETEP on pieces of DNA with 64 processors. The total time taken by each DNA piece is plotted as a function of the number of atoms. Also shown are times for calculations of equivalent quality with CASTEP. More details on the ONETEP calculations are shown in the inset: on the left axis the number of NGWF iterations is plotted as a function of the number of atoms (triangles) and on the right axis the time per iteration in hours is plotted as a function of the number of atoms (squares).

number of atoms for a fixed number of processors. We have used DNA fragments of increasing length to test the linear-scaling properties of ONETEP with the number of atoms. The structures are of *B*-DNA which is the form in which DNA is commonly encountered in physiological conditions within cells and are constructed by repeating an alternating sequence of adenine-thymine and guanine-cytosine base pairs. We have used an orthorhombic simulation cell for these calculations with dimensions $30\ \text{Å} \times 30\ \text{Å} \times 220\ \text{Å}$. It is possible to have such massive simulation cells in our calculations because in contrast to the conventional plane wave approach where empty space is very expensive in memory, ONETEP involves only atom-localized quantities and empty space costs little. While our simulation cell obeys periodic boundary conditions, it is so large that the *supercell approximation*[46] holds extremely well, i.e., all our DNA pieces are nonperiodic (their chemically active ends were terminated by hydrogen atoms) and are so far apart from their periodic images that for all intents and purposes they can be considered as isolated. The DNA calculations were performed at a psinc grid spacing equivalent to a plane wave kinetic energy cutoff of 42 Ry. We have used standard norm-conserving pseudopotentials for all elements taken from the CASTEP (Ref. 50) library of pseudopotentials. The radii of the NGWF localization spheres r_{loc}^{α} were set to 3.2 Å for the hydrogens and 3.3 Å for other elements while the cutoff threshold r_{cut} for the density kernel $\mathbf{K}$ was set to 13.2 Å. These generous thresholds yield results practically indistinguishable from the infinite cutoff limit, yet still the time and memory taken by our calculations increase only linearly with the number of atoms rather than as the cube. We have performed calculations on fragments with 2, 6, 10, 20, 30, and 40 base pairs ranging from 117 to 2606 atoms using 64 processors. Table I summarizes our results. We considered the conjugate gradients optimization of the energy to be converged when the root-mean-square value of the gradient with respect to all NGWFs [Eq. (8)] was less than $10^{-6}\ E_{\text{h}}\ a_0^{3/2}$. This threshold leads to convergence in the energies of $10^{-8}\ E_{\text{h}}$ per atom or better as shown in Table I.

In Fig. 7 we plot the time taken to calculate the total energy for each piece of DNA as a function of the number of atoms. We observe that the curve obtained is essentially a straight line. To compare with a conventional cubic-scaling plane wave code we also show in the same figure calculations with CASTEP,[50] again on 64 processors, with the same kinetic energy cutoff, pseudopotentials, and convergence thresholds. What differs is that we are restricted to using a simulation cell with much smaller dimensions $30\ \text{Å} \times 30\ \text{Å} \times 30\ \text{Å}$ for the CASTEP calculations as the memory requirements for its delocalized orbitals are proportional to the volume of the simulation cell. The largest piece of DNA that can fit in this simulation cell is only eight base pairs long (510 atoms) but as we can see in Fig. 7 the cost of the calculation due to cubic scaling is already so severe that, even without the memory limitations, adding more atoms would soon lead to unfeasibly long computing times. The inset in Fig. 7 focuses on two important points about the ONETEP calculation. First, the cost of each iteration is indeed linear with the num-

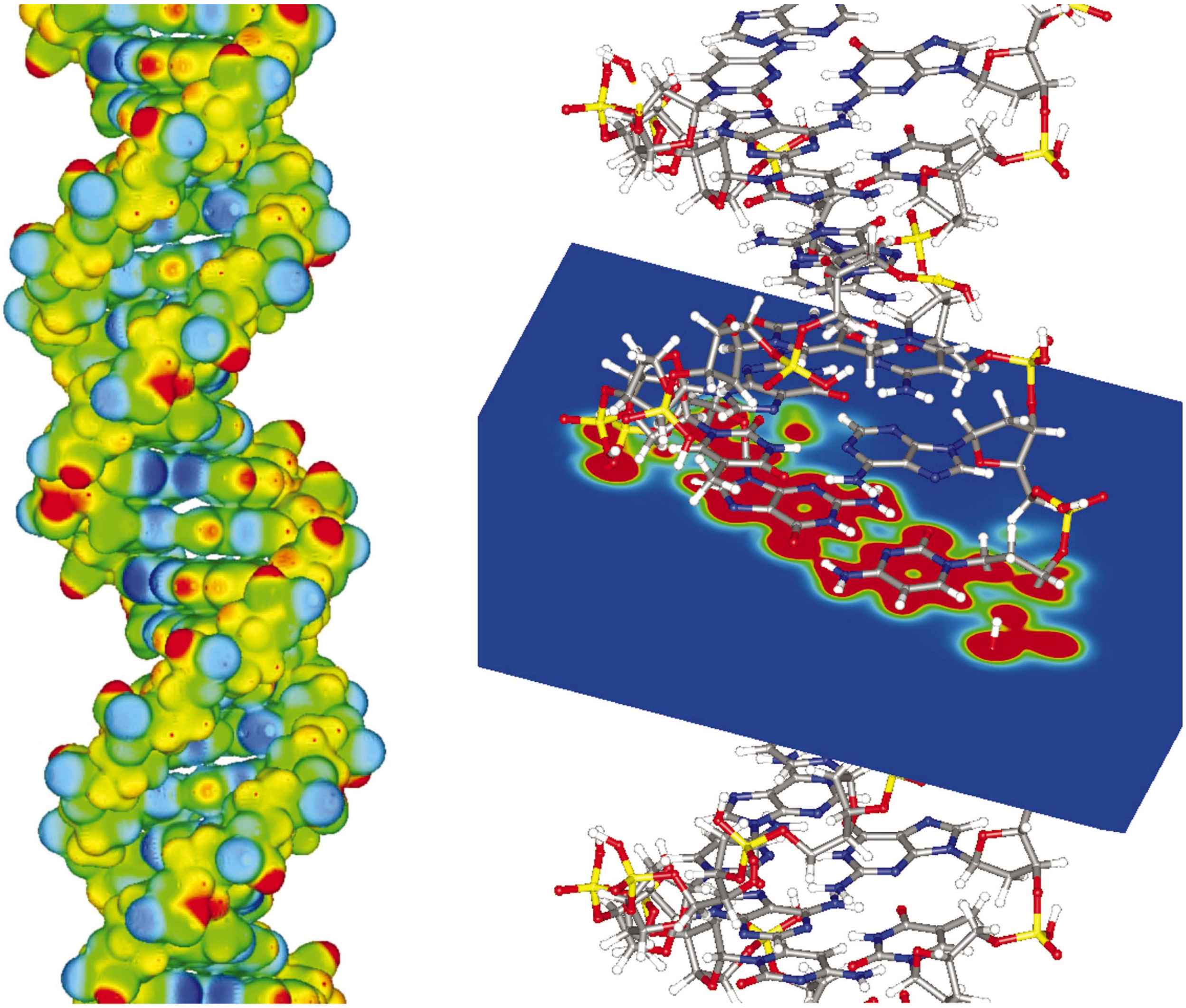

FIG. 8. (Color) ONETEP calculations on a 20 base-pair piece of DNA (1296 atoms). Left, electrostatic potential plotted on an isosurface of the charge density. Right, charge density contours on the plane of the hydrogen bonds of a cytosine-guanine pair.

ber of atoms. Second, the linearity of the total computational time as a function of the number of atoms is a result of our preconditioning scheme[48] which ensures this ‘‘true linear-scaling’’ behavior by making the number of iterations small and independent of the number of atoms, as mentioned in Sec. III.

Quantum mechanical calculations produce a great deal of information that can be difficult to analyze, especially when very large numbers of atoms are involved as in the case of ONETEP. It is therefore very helpful to have the capability to extract information in a visual manner. For this reason we have built into ONETEP the functionality to output information from our calculations in a form suitable for visualization. An example is shown in Fig. 8 where we show three dimensional plots obtained from our calculation on the 20 base-pair (1296 atom) piece of DNA. On the left side of Fig. 8 we show an isosurface of the charge density corresponding to a value of $0.005\,e^{-}/a_0^3$ which is a quantity directly comparable with experimental x-ray diffraction data. The shape of this surface is very informative: one can distinguish a major and a minor groove which are characteristic of the structure of *B*-DNA. It also gives us an indication of the areas most likely to be reached by enzymes intended to dock with DNA. This isosurface is colored according to the values of the total electrostatic potential, ranging from blue for the low values to red for the high values. Another useful plot is shown on the right side of Fig. 8 which depicts contours of the charge density on the plane defined by the heterocyclic rings of a cytosine-guanine base pair. These contours clearly show that the bases are connected with three hydrogen bonds and their relative strengths are also indicated.

Due to the relationship of ONETEP with the conventional plane wave approach, we can often take advantage of the significant technical experience which has been accumulated when adding functionality to the code. Thus, we have already implemented in ONETEP a range of well established GGA exchange-correlation functionals. We show in Table II how our calculations with these functionals compare with the well-established CASTEP code. As a test system we used the smallest of our DNA pieces (two base pairs, 117 atoms, its structure is shown on the left of Fig. 7) and a much smaller simulation cell (dimensions 20 Å×20 Å×20 Å) so that the CASTEP calculations do not take excessive amounts of time to run. All the other parameters were kept the same as in our

TABLE II. The binding energy in kcal/mol for the two strands of a two base-pair piece of DNA (117 atoms) as calculated with various exchange-correlation functionals with ONETEP (this work) and with CASTEP (Ref. 50). The % difference of the ONETEP results with respect to CASTEP is also shown.

Functional	ONETEP	CASTEP	Difference (%)
LDA[a,b]	61.7	61.4	0.5
PBE[c]	42.6	42.3	0.7
RPBE[d]	32.9	32.4	1.5
PW91[e]	44.0	43.5	1.1

[a] Reference 62.
[b] Reference 63.
[c] Reference 64.
[d] Reference 65.
[e] Reference 66.

larger DNA calculations. The quantity we compare here is the binding energy between the DNA piece and its two isolated strands. This binding energy mainly arises from the five hydrogen bonds (two from the adenine-thymine pair and three from the guanine-cytosine pair) that keep the DNA together. In Table II we show the binding energies we obtained for the various functionals with ONETEP and CASTEP. We observe that the agreement between the two codes is exceptionally good with differences in the range 0.5%–1.5%.

VI. CONCLUSIONS

We have presented ONETEP, an implementation for parallel computers of our linear-scaling DFT method[27] which is a reformulation of the conventional cubic-scaling plane wave approach. We have shown that by exploiting the real space localization of the electronic system that is inherent in nonmetallic materials, we are able to optimize with linear cost strictly localized quantities expressed in terms of a delocalized plane wave basis. These same localized quantities have led us to a physical way of dividing the computational effort among many processors to achieve excellent computational speedups with increasing numbers of processors. We have confirmed that the time taken by ONETEP as a function of increasing number of atoms for a given number of processors is indeed linear, which means that we can take full advantage of computational resources. We have performed density functional calculations on a number of systems containing thousands of atoms which confirm that the localization is always achieved in a controlled and mathematically consistent manner. Thus, ONETEP provides the same accuracy as conventional cubic-scaling plane wave approaches, and offers fast and stable convergence. We believe that ONETEP will open the door to a whole new level of accurate large scale simulation with enormous potential for applications in the computational modeling of problems in important areas such as nanoscience and biophysics. We will soon be able to use ONETEP to optimize structures and perform dynamical simulations. Furthermore we expect that the common starting point we share with the conventional plane wave approach will facilitate the reformulation to the ONETEP framework of all the computational machinery that has been developed for the calculation of important experimental observables such as second and higher order derivatives of the energy to external perturbations,[58] nuclear magnetic resonance chemical shifts,[59] or changes in electric polarization.[60] By coupling ONETEP with a new hybrid scheme[61] for classical mechanical simulations with quantum accuracy in required regions we can also envisage a capability to perform simulations with millions of atoms and thus approach problems well into the mesoscopic regime.

ACKNOWLEDGMENTS

One of the authors (C.-K.S.) would like to thank Professor Volker Heine for his constant support throughout this work and for many helpful discussions. The authors would like to thank Dr. Lucio Colombi Ciacchi for providing them with the atomic coordinates of a structure of DNA. C.-K.S. and A.A.M would like to thank the EPSRC (Portfolio Partnership Grant No. GR/S61263/01) for postdoctoral research funding. P.D.H. would like to thank Sidney Sussex College, Cambridge for a research fellowship. Crucial to this work was the ready access to the ''Franklin'' SUN Fire supercomputer of the Cambridge-Cranfield high performance computing facility (CCHPCF) for which the authors are grateful.

[1] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, 1st ed. (McGraw-Hill, New York, 1989).
[2] R. M. Martin, *Electronic Structure. Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, 2004).
[3] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[4] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[5] C.-K. Skylaris, O. Igglessi-Markopoulou, A. Detsi, and J. Markopoulos, Chem. Phys. **293**, 355 (2003).
[6] C. Molteni, I. Frank, and Parrinello, J. Am. Chem. Soc. **121**, 12177 (1999).
[7] E. Artacho, M. Rohlfing, M. Côté, P. D. Haynes, R. J. Needs, and C. Molteni, Phys. Rev. Lett. **93**, 116401 (2004).
[8] A. H. Nevidomskyy, G. Csányi, and M. C. Payne, Phys. Rev. Lett. **91**, 105502 (2003).
[9] L. Colombi Ciacchi and M. C. Payne, Phys. Rev. Lett. **92**, 176104 (2004).
[10] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).
[11] G. Galli, Curr. Opin. Solid State Mater. Sci. **1**, 864 (1996).
[12] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
[13] W. Kohn, Phys. Rev. **115**, 809 (1959).
[14] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
[15] E. I. Blount, Solid State Phys. **13**, 305 (1962).
[16] N. Marzari and D. Vanderbilt, Phys. Rev. B **56**, 12847 (1997).
[17] C. M. Goringe, E. Hernández, M. J. Gillan, and I. J. Bush, Comput. Phys. Commun. **102**, 1 (1997).
[18] M. Challacombe, Comput. Phys. Commun. **128**, 93 (2000).
[19] J.-L. Fattebert and J. Bernholc, Phys. Rev. B **62**, 1713 (2000).
[20] C. K. Gan and M. Challacombe, J. Chem. Phys. **118**, 9128 (2003).
[21] E. R. Davidson and D. Feller, Chem. Rev. (Washington, D.C.) **86**, 681 (1986).
[22] C. F. Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends, Theor. Chem. Acc. **99**, 391 (1998).
[23] P. D. Haynes and M. C. Payne, Comput. Phys. Commun. **102**, 17 (1997).
[24] J. Junquera, O. Paz, D. Sánchez-Portal, and E. Artacho, Phys. Rev. B **64**, 235111 (2001).
[25] E. Hernández, M. J. Gillan, and C. M. Goringe, Phys. Rev. B **55**, 13485 (1997).
[26] P. Fernández, A. Dal Corso, A. Baldereschi, and F. Mauri, Phys. Rev. B **55**, R1909 (1997).
[27] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, and M. C. Payne, Phys. Rev. B **66**, 035119 (2002).
[28] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, C. J. Pickard, and M. C. Payne, Comput. Phys. Commun. **140**, 315 (2001).
[29] C.-K. Skylaris, O. Diéguez, P. D. Haynes, and M. C. Payne, Phys. Rev. B **66**, 073103 (2002).
[30] R. Baer and M. Head-Gordon, Phys. Rev. Lett. **79**, 3962 (1997).

[31] S. Ismail-Beigi and T. A. Arias, Phys. Rev. Lett. **82**, 2127 (1999).
[32] L. He and D. Vanderbilt, Phys. Rev. Lett. **86**, 5341 (2001).
[33] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[34] C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **253**, 268 (1996).
[35] M. C. Strain, G. E. Scuseria, and M. J. Frisch, Science **271**, 51 (1996).
[36] E. Schwegler and M. Challacombe, J. Chem. Phys. **105**, 2726 (1996).
[37] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002).
[38] X. P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).
[39] J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997).
[40] P. D. Haynes and M. C. Payne, Phys. Rev. B **59**, 12173 (1999).
[41] U. Stephan, Phys. Rev. B **62**, 16412 (2000).
[42] J. Kim, F. Mauri, and G. Galli, Phys. Rev. B **52**, 1640 (1995).
[43] O. F. Sankey and D. J. Niklewski, Phys. Rev. B **40**, 3979 (1989).
[44] S. D. Kenny, A. P. Horsfield, and H. Fujitani, Phys. Rev. B **62**, 4899 (2000).
[45] E. Anglada, J. M. Soler, J. Junquera, and E. Artacho, Phys. Rev. B **66**, 205101 (2002).
[46] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).
[47] In our earlier work (Refs. 27 and 54) we refer to these functions as ‘‘periodic bandwidth limited delta functions.’’
[48] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, and M. C. Payne, J. Chem. Phys. **119**, 8842 (2003).
[49] E. Artacho and L. M. del Bosch, Phys. Rev. A **43**, 5770 (1991).
[50] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne, J. Phys.: Condens. Matter **14**, 2717 (2002).
[51] T. P. Straatsma, E. Apra, T. L. Windus *et al.*, NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.5, Pacific Northwest National Laboratory, Richland, Washington, 2003.
[52] T. H. Dunning, Jr., J. Chem. Phys. **90**, 1007 (1989).
[53] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, J. Chem. Phys. **96**, 6796 (1992).
[54] A. A. Mostofi, C.-K. Skylaris, P. D. Haynes, and M. C. Payne, Comput. Phys. Commun. **147**, 788 (2002).
[55] P. Pulay, Mol. Phys. **17**, 197 (1969).
[56] *Message Passing Interface Forum*, http://www.mpi-forum.org/
[57] P. Pacheco, *Parallel Programming with MPI* (Morgan Kaufmann, San Fransisco, CA, 1996).
[58] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).
[59] C. J. Pickard and F. Mauri, Phys. Rev. B **63**, 245101 (2001).
[60] R. Resta, Int. J. Quantum Chem. **75**, 599 (1999).
[61] G. Csányi, T. Albaret, M. C. Payne, and A. De Vita, Phys. Rev. Lett. **93**, 175503 (2004).
[62] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[63] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[64] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[65] B. Hammer, L. B. Hansen, and J. K. Nørskov, Phys. Rev. B **59**, 7413 (1999).
[66] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Phys. Rev. B **46**, 6671 (1992).