PHYSICAL REVIEW B VOLUME 51, NUMBER 3 15 JANUARY 1995-I

Linear system-size scaling methods for electronic-structure calculations

Pablo Ordejón
Department of Physics and Materials Research Laboratory, University of Illinois, Urbana, Illinois 61801

David A. Drabold
Department of Physics and Astronomy, Ohio University, Athens, Ohio 45701-2979

Richard M. Martin
Department of Physics and Materials Research Laboratory, University of Illinois, Urbana, Illinois 61801

Matthew P. Grumbach
Department of Physics and Astronomy and Materials Research Group in High Pressure Materials Synthesis, Box 871504, Arizona State University, Tempe, Arizona 85287-1504
(Received 18 May 1994)

We describe a method for performing electronic-structure calculations of the total energy and interatomic forces which scales linearly with system size. An energy functional is introduced which possesses a global minimum for which (1) electronic wave functions are orthonormal and (2) the correct electronic ground-state energy is obtained. Linear scaling is then obtained by introducing a spatially truncated Wannier-like representation for the electronic states. The effects of this representation are studied in detail. Molecular-dynamics simulations using an orthogonal tight-binding basis and *ab initio* local-orbital density-functional methods are presented. We study both Car-Parrinello and conjugate-gradient molecular-dynamics schemes and discuss practical methods for dynamical simulation. A detailed connection between our method and the density matrix approach of Daw [Phys. Rev. B **47**, 10 895 (1993)] and Li, Nunes, and Vanderbilt [Phys. Rev. B **47**, 10 891 (1993)] is also provided.

I. INTRODUCTION

Recent advances in computational methods have made possible first-principles calculations for materials far beyond anything possible only a few years ago. The work of Car and Parrinello,[1] in particular, initiated an entire new set of methods[2] which can be applied to any independent-particle theory such as the density-functional approach. Current methods have in common that the total energy is written as a functional of the electronic states and efficient techniques are used to iterate the wave functions toward the minimum energy solution. Although they are vastly more efficient than traditional matrix diagonalization methods (especially for large basis sets, like plane waves), the most widely used methods still have the same scaling as diagonalization—for large systems the computational time increases asymptotically as the cube of the number of electrons considered. In fact, current algorithms intrinsically involve a nonlinear dependence upon the number of electrons, since they require that each orbital be orthogonal to the others, which in general would appear to scale at least as the square of the number of electrons.

On the other hand, it is clear that this nonlinear scaling is not inescapable. It is well known that the local properties of a region can be computed from knowledge of the electronic states only in some vicinity of that region. This is established, for example, by expressing properties in terms of Green's functions or density matrices. This has been emphasized particularly by Heine and co-workers.[3] One of the methods they have applied very successfully is the recursion method,[4] which is a continued fraction representation of an expansion in coordination shells of interactions. The continued fraction can be terminated in a range that allows feasible calculations. Quantum mechanics requires that the properties at one point cannot be purely local; however, the effects decay rapidly (exponentially in the case of insulators) and can be considered to vanish outside some range of nonlocality. Clearly, for large enough systems (larger than the range of the nonlocality), calculations by this method are linear in the size of the system, since different parts can be computed independently. Many other workers have employed methods which are based upon generation of Green's functions or an equivalent representation.[5–12]

The aim of the present work is to develop a method that will combine the virtues of these two different approaches, that is, a method which can scale linearly with system size for large systems, and yet utilize the computational advances of the Car-Parrinello-type methods. One way to accomplish this was suggested by Galli and Parrinello,[13] who pointed out that Car-Parrinello-type methods can be modified to scale linearly for large systems if the electronic orbitals are required to be localized, Wannier-like functions. For purposes of determining integrated properties such as the total energy or forces on

the atoms, it is always possible to transform to a set of orbitals which are localized. In a metal they have power-law localization and in an insulator they are exponentially localized.[14] If one can neglect the overlap of these functions beyond some range, then it is in principle possible to construct an algorithm which scales linearly with the system size, since a given electron orbital overlaps only a finite number of other orbitals, independent of the size of the system.

Several groups have independently proposed methods to make practical algorithms based upon localized Wannier-like functions[15–21] or localized orbitals.[22] Here we describe in detail one method.[15] The algorithm generated clearly scales linearly in principle; however, for reasons that will be explained below, the constraint requiring the wave functions to be strictly localized can lead to difficulties in applying efficient iterative methods. To show that the methods can be practically implemented, an important part of the present paper is the demonstration that these difficulties can be overcome in realistic cases.

Here we will emphasize application of the linear scaling methods in cases where one wishes to calculate properties of an entire system, for example, as would happen in a thermal simulation or a global relaxation. This is in contrast to the spirit of the Green's function methods, which emphasize that one can calculate the properties of part of a large system without ever dealing with the entire system. Such an approach seems preferable if one seeks only a local relaxation, for example. We note that the present methods can also be adapted to work in this way as well—the only change required is to allow variations in the wave functions only in a localized region, but enforcing orthogonality to a finite number of surrounding fixed localized functions. While this is not of primary concern in this paper, we show in Sec. VII that the calculation of the force constant matrix is particularly efficient since it may be built by considering small *local* displacements of each of the atoms. This provides an efficient means of computing the vibrational power spectrum.

While the present method does not explicitly produce eigenvalues, we note that existing techniques allow the computation of selected eigenvalues or the density of states in an order-N fashion.[23–25] This makes it easy to compute the optical gap, for instance. These methods can be used as a complement to the present technique.

In the process of generating the present algorithm, we found it very convenient to express our minimization procedure in such a way that orthogonalization is not explicitly performed. We will define a functional which when freely minimized with no constraints leads to the correct energy and orthogonal functions. We believe this procedure has general utility in many different methods, whether linearly scaling or not. For instance, it has proven to be useful in cases where a self-consistent solution cannot be found by standard techniques due to the instability of the solution, like the one-dimensional Hubbard model.[26] The basic idea is that a minimization procedure which involves no constraints is more robust and may be expected to better utilize minimization algorithms which are available. In this respect our work is quite close to that of Mauri, Galli, and Car[17,18] and also has similarities to work of Wang and Teter.[16] There are also strong links between our orbital formulation and the density matrix approach of Li *et al.*[27] and Daw.[28] The localization properties of the density matrix were exploited by Yang[29] to produce an order-N algorithm.

The method presented here also represents a way to calculate approximate Wannier functions, which may have many possible uses. For example, Nunes and Vanderbilt[30] have recently pointed out a way to calculate the dielectric response of insulators using this approach. This is a solution to a problem which is quite difficult in the usual representation of extended Bloch states.

This paper is organized as follows. In Sec. II we discuss general aspects of solving the noninteracting-electron problem, to put the present work in context. In Sec. III we describe the use of localized Wannier-like functions to represent the electronic states. In Sec. IV we give a detailed analysis of an energy functional which may be directly minimized without any subsidiary orthonormality condition. When combined with the Wannier picture of electronic states, linear system-size scaling is obtained. In Sec. V, the method is implemented with a simple tight-binding Hamiltonian to illustrate the practical implementation of the scheme. The only important approximation of the method, the truncation of the Wannier-like states, is discussed in Sec. VI with some elementary but revealing examples of its effects. Quantitative examples of these effects for practical problems are given in Sec. VII, whereas in Sec. VIII we describe means for performing accurate molecular-dynamics simulations. Section IX contains the conclusions of our work. Additionally, in Appendix A we provide some rigorous properties of the new energy functional, and in Appendix B a detailed connection of the present work with the density matrix method proposed by Li *et al.*[27] and Daw[28] is given.

II. BACKGROUND

We will consider a system of N noninteracting electrons of each spin, moving in an effective potential $V_{\rm eff}(\mathbf{r})$, so that the effective one-particle Hamiltonian is

$$H_{\rm eff}(\mathbf{r}) = -\frac{\hbar^2}{2m}\nabla^2 + V_{\rm eff}(\mathbf{r}) \ . \tag{1}$$

The effective potential $V_{\rm eff}(\mathbf{r})$ includes the Coulomb interaction of the electrons with the nuclei and any external potential and must also simulate the many-body electron-electron interactions in an effective way. The detailed form of the effective one-electron Hamiltonian depends on the particular approximation that one uses. In empirical tight binding,[31] the matrix elements $h_{\mu\nu}$ of the Hamiltonian in an atomiclike orbital basis $\{|\phi_\mu\rangle\}$ are assumed to be known and in practice are adjusted to fit experimental or theoretical information such as the band structure. On the other hand, in *ab initio* approaches such as Hartree-Fock[32] or density functional[33] (DF) in the local-density approximation[34] (LDA), the interactions are obtained from first principles, making

some approximations to map the many-body problem of the interacting electrons onto an effective system of noninteracting particles. In the Hartree-Fock method, the effective potential contains the Coulomb repulsion of each electron with the electronic charge created by the rest of the electrons (Hartree term) and a nonlocal exchange term that arises from the correlated motion of the electrons due to the Pauli principle. In the Kohn-Sham formulation of density-functional theory, the effective Hamiltonian takes the form

$$H_{\mathrm{LDA}}(\mathbf{r}) = -\frac{\hbar^2}{2m}\nabla^2 + V_{\mathrm{ion}}(\mathbf{r}) + V_H(\mathbf{r}) + V_{\mathrm{XC}}(n(\mathbf{r})) \ , \tag{2}$$

where V_{ion} is the Coulomb potential of the nuclei situated at positions $\{\mathbf{R}_l\}$, V_H is the Hartree potential of the electrons

$$V_H(\mathbf{r}) = e^2 \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d^3r' , \tag{3}$$

and V_{XC} is the exchange-correlation potential. In the local-density approximation $V_{\mathrm{XC}}(\mathbf{r})$ depends on the value of the charge density $n(\mathbf{r})$. The LDA Hamiltonian can also include a nonlocal pseudopotential term to eliminate the core electrons from the calculation.

For the noninteracting electrons that we are considering, the state of the system can be described specifying N one-electron states $\{|\chi_i\rangle\}$, $i = 1, ..., N$, each of which is occupied with two electrons of opposite spin. $\{|\chi_i\rangle\}$ is one of the possible bases of the *occupied space* ($\mathcal{H}_{\mathrm{occ}}$), which is a subspace of $\mathcal{H}$, the Hilbert space of the system. Physical properties such as the charge density and the electronic energy can be obtained within this occupied space without any need to make reference to the rest of the Hilbert space. Although the most usual procedure is to use orthonormal sets of states to describe the occupied space, this is not a fundamental limitation, and one can also use nonorthonormal sets to describe the quantum system. In this case, the usual matrix representation for orthogonal basis generalizes to a tensorial representation for the vectors and operators in the Hilbert space, as shown by Artacho and Miláns del Bosch.[35] If the set of states defining the occupied space are nonorthogonal, the overlap matrix

$$S_{ij} = \langle \chi_i | \chi_j \rangle \tag{4}$$

is not the identity matrix δ_{ij}. In this case, the density matrix operator, defined as the projection onto the occupied space, is given by the expression

$$\hat{\rho} = 2 \sum_{i,j=1}^{N} |\chi_i\rangle (S^{-1})_{ij} \langle \chi_j| \ , \tag{5}$$

where $(S^{-1})_{ij}$ is the element i,j of the inverse of the S matrix, and the factor 2 is for spin. The charge density is simply the diagonal elements of $\hat{\rho}$ in the coordinate representation

$$\rho(\mathbf{r}) = 2 \sum_{i,j=1}^{N} \chi_i(\mathbf{r}) (S^{-1})_{ij} \chi_j^*(\mathbf{r}). \tag{6}$$

The calculation of the electronic total energy depends on the specific one-electron model used. There is, however, a term common to all the approximations: the band structure energy, which is defined as the trace of the Hamiltonian in the occupied space. In terms of the matrix elements of the Hamiltonian $\hat{H}$ in the occupied space basis $\{|\chi_i\rangle\}$, the band structure energy takes the form

$$\begin{aligned} E_{\mathrm{BS}} = 2\mathrm{Tr}(\hat{H}) &= 2 \sum_{i,j=1}^{N} (S^{-1})_{ij} H_{ji} \\ &= 2 \sum_{i,j=1}^{N} (S^{-1})_{ij} \langle \chi_j | \hat{H} | \chi_i \rangle \ . \end{aligned} \tag{7}$$

In empirical tight binding,[31] the total electronic energy is usually taken as the band structure energy plus a repulsive two-body potential V_R between the atoms that is fitted to experiment together with the electronic interactions $h_{\mu\nu}$:

$$E = E_{\mathrm{BS}} + \frac{1}{2}{\sum_{l,l'}}' V_R(|\mathbf{R}_l - \mathbf{R}_{l'}|) \ . \tag{8}$$

In the LDA, the total electronic energy is obtained by adding to the band structure energy two terms which account for double counting in the Coulomb electronic repulsion and for exchange-correlation corrections:

$$\begin{aligned} E = E_{\mathrm{BS}} + \frac{e^2}{2} \Bigg[&- \int n(\mathbf{r}) d^3r \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d^3r' \\ &+ \int n(\mathbf{r}) \left[\epsilon_{\mathrm{XC}}(n) - V_{\mathrm{XC}}(n) \right] d^3r \Bigg] \ . \end{aligned} \tag{9}$$

These correction terms can be calculated using Eq. (6) for the charge density, even for the case of nonorthonormal states $\{|\chi_i\rangle\}$.

The ground state of the system is the one that minimizes the energy. There are two "standard" procedures to find this ground state: direct diagonalization and iterative minimization. In direct diagonalization one diagonalizes the whole Hamiltonian matrix (which has dimension $M \times M$ if the basis set has M elements) and obtains all the eigenvectors. The band structure energy and the charge density are computed using the energies and wave functions of the N lowest eigenstates. In the Hartree-Fock or Kohn-Sham LDA methods, the procedure is repeated until self-consistency is achieved. For non-self-consistent Hamiltonians, one diagonalization is sufficient. In iterative minimization the total energy is minimized with respect to variations in the states $\{|\chi_i\rangle\}$. Although some authors[20,21,36,37] have developed schemes to deal with nonorthonormal sets of states $\{|\chi_i\rangle\}$, calculating the inverse of the overlap matrix (S^{-1}) and using Eq. (7), the most common procedure is to restrict the minimization to orthonormal states, imposing explicitly the condition

$$S_{ij} = \langle \chi_i | \chi_j \rangle = \delta_{ij} \tag{10}$$

by means of an orthonormalization procedure (usually Gram-Schmidt) during the energy minimization. The

advantage of this extra step is that the band structure energy and the charge density take a simple form:

$$E_{\rm BS} = 2\sum_{i=1}^{N} H_{ii} \ , \tag{11}$$

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N} |\chi_i(\mathbf{r})|^2, \tag{12}$$

avoiding the calculation of the S^{-1} matrix and the appearance of numerical instabilities. The iterative minimization techniques are very convenient for calculations involving large systems or large basis sets, since only the occupied states are calculated and stored. For non-self-consistent Hamiltonians, the true ground-state band energy is given by the minimum value of Eq. (11) in the space of all functions $\{|\chi_i\rangle\}, i = 1, ..., N$, which satisfy the constraints Eq. (10). For self-consistent theories, the solution is found by minimizing the total energy [Eq. (9) for the LDA], using Eq. (11) for the band structure energy, updating the charge density and the Hamiltonian at the same time, so that minimization and self-consistency are simultaneously achieved. For a review of *ab initio* iterative minimization techniques we refer the reader to Ref. 2.

One limitation of these traditional methods is their unfavorable scaling with system size. For large enough systems the computational time needed increases as N^3, due to the necessary orthogonalization step. On the other hand, the memory required to store the wave functions scales as N^2, and it has been argued[38] that this is the real limit for the size of systems that can be handled by currently available computational platforms. In any case, it is clear that a method with linear scaling, both in the computation time and in the memory requirements, would be of great value for the study of large systems which cannot be treated with the standard approaches. We describe one such method in the following sections.

III. LOCALIZED WAVE FUNCTIONS

Several authors have developed linear scaling algorithms making use of localized Wannier-like functions. These methods explicitly exploit the fact that the set of states $\{|\chi_i\rangle\}$ that defines the ground state of the system is not unique, since any unitary transformation of the N lowest eigenvectors $\{|\psi_i\rangle\}$ has the same energy. In particular, we can describe the ground state of the system by means of orthonormal localized wave functions (LWF's) centered on different points. In the case of a periodic system, these LWF's are simply the Wannier functions. Although, strictly speaking, the LWF's are spread throughout the system, they decay very rapidly with the distance from the center of localization. This decay is exponential in insulators and a power law in metals. Kohn[14] has pointed out that, for an insulator with gap E_g, the best localized Wannier functions decay exponentially with a rate $\kappa \sim (2mE_g/\hbar^2)^{1/2}$, where m is the electron mass. For realistic systems, Wannier functions have been computed by a number of authors.[39] As an illustration, in Fig. 1 we show a Wannier function for a one-dimensional (1D) tight-binding chain at half filling. The chain has two alternating kinds of orthogonal s orbitals with a difference in on-site energies of 1 eV, which interact with a first-neighbor hopping $V = -1$ eV. Figure 1 shows one of the Wannier functions of the system, the one which is centered on atom number 49. Since the system has an energy gap of 1 eV, the decay of the LWF is exponential. Therefore, although in order to describe the system *exactly* with LWF's one should allow them to spread through all the system, we can, as an approximation, truncate the LWF's beyond a certain cutoff radius R_c from the center. Note that the typical localization range does not depend essentially on the size of the system but on the nature of the bonding and the electronic states (metallic or insulating), and the energy gap. In this case, truncating the localized wave functions beyond R_c would produce an order-N scaling. Indeed, if the LWF's are zero outside the localization range, the Hamiltonian and overlap matrix elements between LWF's centered on distant points will be zero, and therefore, for each localized orbital $|\chi_i\rangle$, only the matrix elements H_{ij} and S_{ij} with orbitals $|\chi_j\rangle$ within a distance of the order of $2R_c$ must be calculated (a number that is independent of the size of the system).

Once we have taken the approach of building localized functions for the solution of the electronic problem, we have to devise an algorithm for the actual calculation of the LWF's and the minimization of the electronic energy. The naive approach would be to use the standard iterative minimization techniques for the minimization of the electronic energy. In principle, it would be convenient to use orthonormal LWF's, since this allows us to use Eq. (11) for the electronic energy, avoiding the calculation of the S^{-1} matrix. This is convenient because, even when the states are strictly localized and the overlap matrix S is sparse, the *exact* S^{-1} matrix is not, and this would be a difficulty for achieving the linear scaling. One possibility is to use an orthonormalization procedure combined with

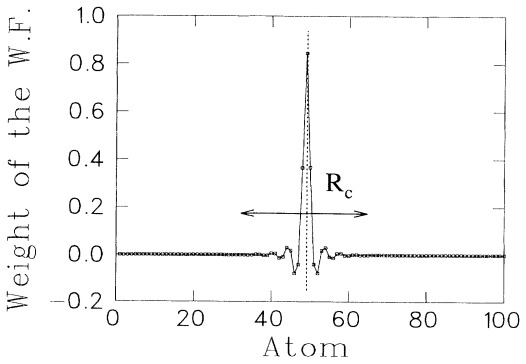

FIG. 1. Wannier function for the 1D chain described in the text. We plot the weight of the Wannier function on each of the atoms in the chain. R_c represents the range beyond which the Wannier function could be truncated without loss of accuracy.

the minimization of the electronic energy Eq. (11) (valid for orthonormal states). The ideal candidate would be the Löwdin orthonormalization scheme,[40] in which, given an initial set of normalized but nonorthogonal wave functions $\{|\chi_i\rangle\}$, a new set $\{|\chi'_i\rangle\}$ is generated by

$$|\chi'_i\rangle = |\chi_i\rangle - \frac{1}{2}\sum_{j\neq i} S_{ij}|\chi_j\rangle, \tag{13}$$

and this is repeated until the wave functions are orthogonal to the desired degree of accuracy. If the initial wave functions are localized, this orthonormalization procedure produces orthonormal functions which are also localized, since in each iteration the functions are only mixed with those in the neighborhood (the ones which have a nonzero overlap). However, each iteration of the orthogonalization increases the localization range. If we want to impose a strict cutoff on the LWF's, we must modify slightly the algorithm to avoid mixing the wave function with components of others outside the localization range. This results in a decrease in the efficiency that will ultimately preclude its practical application with realistic values of the localization range. To illustrate this point, we have performed a calculation using this scheme in a 64-atom Si supercell in the diamond structure, with the atoms displaced 0.05 Å in the direction of the LO phonon mode. We have used an *ab initio* tight-binding Hamiltonian[41] with an sp^3 basis on each Si atom, with interactions and overlaps up to third nearest neighbors. The localized wave functions are centered on the covalent bonds, and we have imposed a cutoff radius of 4.0 Å. In Fig. 2 we show the maximum deviation from orthonormality (i.e., the maximum value of $|S_{ij}|$) versus the number of iterations in the orthonormalization procedure. We observe that, although the value of the maximum overlap decreases rapidly in the first iterations, the process degrades after a while, and the orthogonality cannot be improved beyond a certain degree of accuracy. This difficulty in imposing a strict orthogonalization to the LWF's is common to other possible orthonormalization schemes that we have tested. Since the calculation of the electronic energy by means of Eq. (11) requires orthonormal wave functions, the errors in the orthonormalization will translate into errors in the values of the energy, which will be more severe for larger deviations from orthogonality. In Fig. 3 we show a steepest descent energy minimization of the electronic energy using the described Löwdin orthogonalization. The results correspond to two different values (0.032 and 0.01) for the tolerance in the orthogonalization procedure. The exact value of the energy, computed with diagonalization, is -1069.32 eV. In the case of the large tolerance the error is so large that the energy oscillates widely and finally diverges. Even for the smaller tolerance the errors are large enough to prevent an accurate determination of the minimum energy. Imposing a stricter tolerance is impossible, as was shown in Fig. 2.

The orthogonalization problem remains, therefore, a serious obstacle to the achievement of linear scaling. When the wave functions are kept localized, the orthogonalization is no longer order N^3, but it is not accurate

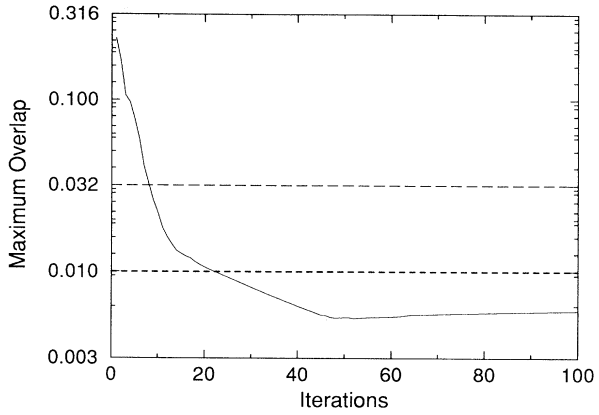

FIG. 2. Maximum deviation from orthonormality vs number of iterations in the Löwdin orthogonalization. The full line corresponds to a typical orthogonalization of strictly localized wave functions in silicon. The cutoff range was 4.0 Å. The dashed lines show the tolerances used in the energy minimizations shown in Fig. 3.

enough to allow the minimization of the electronic energy. We must, therefore, find a way to avoid the orthogonalization in order to find a reliable order-N algorithm. With that in mind, in the next section we propose an energy functional that, when freely minimized (i.e., with no orthonormalization constraints) leads to the correct ground-state energy and a set of orthogonal wave functions. It will be shown that this functional is well suited to be used with localized wave functions, a combination which produces an efficient order-N algorithm.

IV. THE NEW ENERGY FUNCTIONAL

In this section we will describe the new energy functional that we have derived, and that will enable us to develop an orbital formulation with order-N scaling. In

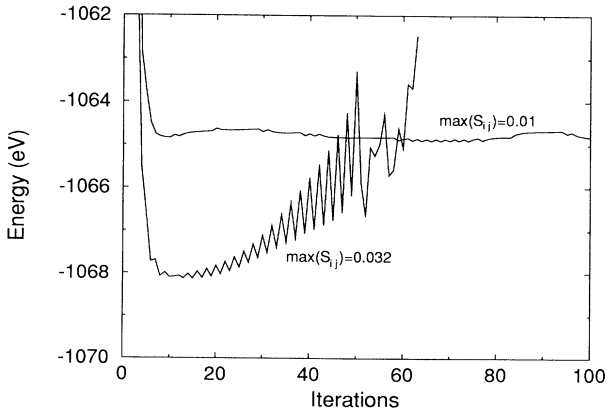

FIG. 3. Energy minimization with localized functions for silicon, using steepest descent and Löwdin orthonormalization. The localized wave functions were centered on the bonds and maintained strictly localized within a radius of 4.0 Å. We show two curves, for different tolerances in the orthogonalization procedure.

our approach, we want to maintain the nice properties of the iterative minimization procedures, but with the purpose of deriving an order-N algorithm. We wish to use a formalism that avoids the calculation of the S^{-1} matrix, since it is nonsparse and can lead to numerical instabilities, and also avoids an explicit orthonormalization step. We therefore want to minimize Eq. (11) (the energy for the case of orthonormal states), *subject to the constraints* that the states are orthonormal (i.e., that $S_{ij} = \delta_{ij}$). The well-known Lagrange multipliers method suggests how to solve this problem. A new functional $\tilde{E}$ is built which consists of the function that has to be minimized plus a term including the constraints multiplied by the Lagrange multipliers (one for each of the constraints):

$$\tilde{E} = 2\left[\sum_{i=1}^{N} H_{ii} - \sum_{i,j=1}^{N} \Lambda_{ji}(S_{ij} - \delta_{ij})\right] . \tag{14}$$

In the usual approach, the Lagrange multipliers Λ_{ji} are considered independent variables, and the solution would be obtained by requiring that $\tilde{E}$ be stationary with respect to variations in both the wave functions $|\chi_i\rangle$ and the Lagrange multipliers Λ_{ji}:

$$\frac{\delta\tilde{E}}{\delta\chi_i^*} = 0 , \tag{15}$$

$$\frac{\delta\tilde{E}}{\delta\Lambda_{ji}} = 0 . \tag{16}$$

Since the second equation implies that the orthonormalization conditions $S_{ij} = \delta_{ij}$ have to be satisfied, solving Eqs. (15) and (16) is equivalent to our original problem. However, these equations suggest an alternative approach. The first of the stationary conditions for $\tilde{E}$, Eq. (15), can be expressed as

$$\frac{\delta\tilde{E}}{\delta\langle\chi_i|} = \hat{H}|\chi_i\rangle - \sum_{j=1}^{N}\Lambda_{ji}|\chi_j\rangle = 0. \tag{17}$$

Taking the product on the left with the vector $\langle\chi_k|$, Eq. (17) leads to

$$H_{ki} - \sum_{j=1}^{N}\Lambda_{ji}S_{kj} = 0 , \tag{18}$$

which, in matrix notation, reads

$$H - S\Lambda = 0 . \tag{19}$$

This equation defines a relation between the Lagrange multipliers and the wave functions $|\chi_i\rangle$ (through the matrix elements of H and S) that must be satisfied at the solution. In that case, the second of the stationary conditions for $\tilde{E}$, Eq. (16), must also hold, so that $S = \mathbb{1}$. Therefore the solution satisfies the equation

$$\Lambda_{ij} = H_{ij} = \langle\chi_i|\hat{H}|\chi_j\rangle. \tag{20}$$

We stress that this relation only holds exactly for the solution, i.e., for the $|\chi_i\rangle$ and Λ_{ij} that make the energy stationary. However, we can *define* a functional with $\Lambda_{ij} = H_{ij}$ even for those χ's which are not in the correct solution. Following this approach leads to the functional

$$\begin{aligned}\tilde{E} &= 2\left[\sum_{i=1}^{N} H_{ii} - \sum_{i,j=1}^{N}(S_{ij} - \delta_{ij})H_{ji}\right] \\ &= 2\mathrm{Tr}\{[\mathbb{1} + (\mathbb{1} - S)]H\}\end{aligned} \tag{21}$$

which is defined for all sets of functions $\{|\chi_i\rangle\}$, which need be neither normalized nor orthogonal. It is apparent that, for a given state defined by the set $\{|\chi_i\rangle\}$, the "modified energy" $\tilde{E}$ will only take the same value as the band energy E_{BS} [Eq. (11)] if it is an orthonormal set, in which case the second term in Eq. (21) will vanish. It is also evident that $\tilde{E}$ will be stationary for those sets $\{|\chi_i\rangle\}$ corresponding to the orthonormal solution, since it has been explicitly built from Eqs. (15) and (16).

In Appendix A we will show that, under certain general conditions, the functional $\tilde{E}$ is not only stationary but has a minimum for the correct orthonormal ground state of the system. Therefore, the *unconstrained minimum* of the functional $\tilde{E}$ coincides with the minimum of E_{BS} *subject to the orthonormality constraints.* We can therefore find the ground-state $\{|\chi_i\rangle\}$ and the ground-state energy E_0 by means of an unconstrained minimization of the functional $\tilde{E}$. The advantage of such a formulation is that no explicit orthonormalization is required, since the second term $\sum_{i,j=1}^{N}(S_{ij} - \delta_{ij})H_{ji}$ in Eq. (21) forces the wave functions toward orthonormality. In general, unconstrained minimization can be performed efficiently by means of standard minimization techniques such as conjugate gradients (CG's). Thus it may be possible to improve upon algorithms which explicitly enforce the constraints. On the other hand, the function $\tilde{E}$ is explicitly nonlinear in the variables $|\chi_i\rangle$, so that the solution may be more complex. We therefore conclude that it is worthwhile to study the direct unconstrained approach, and also to consider possible comparisons of this method with the usual constrained methods[2] and with other approaches which explicitly consider the S^{-1} matrix.[20,21,36]

The functional we have defined is related to the one proposed by Wang and Teter,[16] who also use localized functions and an unconstrained minimization approach. There are, however, two salient differences between the two schemes. (i) Wang and Teter use normalized LWF's and a penalty function to enforce approximate orthogonality. This penalty function is different from the second term of our Eq. (14) in that it contains a single adjustable parameter λ instead of our Lagrange matrix Λ_{ij}, and is built to minimize the overlap between neighbor functions, whereas our functional has no free parameters and simultaneously achieves normality and orthogonality. The parameter λ must be chosen properly, since the results depend critically on its value. (ii) Wang and Teter restrict the LWF's to the orbitals of the two atoms forming each bond, and as a consequence they must reparametrize the tight-binding interactions in order to obtain close agreement with exact diagonalization results. We do not need

to do that, since our energy functional gives an accurate variational approximation to the exact energy when the localization range is large enough (typically two shells of neighbors for silicon). Obviously, our approach requires a larger computational effort, but eliminates any parametrization.

It is interesting to note that a whole family of functionals can be generated in the same spirit as Eq. (21). As we have discussed, Eq. (20) only holds strictly for the orthonormal solution that minimizes the energy Eq. (11), and therefore $S = \mathbb{1}$. If, however, we are near that orthonormal solution, so that $S_{ij} - \delta_{ij}$ are small, we could think that a solution better than $\Lambda_{ij} = H_{ij}$ to Eq. (19) would be obtained if the departure from orthonormality was taken into account. This can be done if we express Eq. (19) as

$$H - \Lambda + (\mathbb{1} - S)\Lambda = 0 \ , \tag{22}$$

so that

$$\Lambda = H + (\mathbb{1} - S)\Lambda \ . \tag{23}$$

We can consider Eq. (23) as a recurrence relationship for Λ so that, given an estimate $\Lambda^{(n)}$, a better one can be obtained:

$$\Lambda^{(n+1)} = H + (\mathbb{1} - S)\Lambda^{(n)}. \tag{24}$$

If we use the exact value of Λ at the solution [Eq. (20)] as a zeroth-order approximation, then the successive approximations to Λ can be written as

$$\begin{aligned} \Lambda^{(n)} &= H + (\mathbb{1} - S)H + \cdots (\mathbb{1} - S)^n H \\ &= \sum_{m=0}^{n} (\mathbb{1} - S)^m H \ . \end{aligned} \tag{25}$$

These approximations of Λ generate a family of functionals $\tilde{E}^{(\mathcal{N})}$ when inserted in Eq. (14):

$$\tilde{E}^{(\mathcal{N})} = 2\mathrm{Tr}\left[\sum_{m=0}^{\mathcal{N}} (\mathbb{1} - S)^m H\right] \ . \tag{26}$$

The functional defined in Eq. (21) is a member of this family, and corresponds to the case $\mathcal{N} = 1$.

The same family of functionals has recently been derived independently by Mauri *et al.*[17] using a different line of reasoning. Instead of using the Lagrange multipliers method to obtain the minimum of the energy Eq. (11) subject to the orthonormality constraints, their starting point is the band structure energy for the general case of nonorthonormal states, Eq. (7). They then replace the inverse of the overlap matrix by its Taylor series expansion in $(\mathbb{1} - S)$ up to an order $\mathcal{N}$:

$$S^{-1} = [\mathbb{1} - (\mathbb{1} - S)]^{-1} \approx \sum_{n=0}^{\mathcal{N}} (\mathbb{1} - S)^n \ , \tag{27}$$

which, when inserted in Eq. (7), produces the family of functionals Eq. (26). In addition, Mauri *et al.* showed that functionals $\tilde{E}^{(\mathcal{N})}$ corresponding to odd-$\mathcal{N}$ have the desired property of having a global minimum at that stable point. Therefore, if we want to be guaranteed to find the solution by means of standard minimization techniques, only the odd-$\mathcal{N}$ functionals are valid candidates.

It is worth noticing that all the odd-$\mathcal{N}$ functionals of this family are equally valid for finding the ground state of the system and that the solution will be independent of which one is used: they will all give an *orthonormal* set of states $\{|\chi_i\rangle\}$ which span the occupied subspace (ground state). However, the numerical performance of the functionals will differ in the number of iterations during the minimization procedure and the number of operations per iteration. It is clear that high-$\mathcal{N}$ functionals will require more operations per iteration, since they require more matrix multiplications to calculate the powers of $(\mathbb{1} - S)$. Besides, as we will discuss in Sec. V, it is crucial to an order-N scheme that the H and S matrices be sparse. Higher-$\mathcal{N}$ functionals will contain higher powers of $(\mathbb{1} - S)$ which will be less and less sparse, eventually destroying the order-N scaling. We therefore have restricted our calculations to the simplest functional $\mathcal{N} = 1$, described by Eq. (21).

Although we will only apply the present method to non-self-consistent Hamiltonians, it can be also used in self-consistent problems, like the Hartree-Fock[26] and LDA methods. There one needs the charge density, which for nonorthonormal states is defined by Eq. (6). The definition consistent with our energy functional, in which S^{-1} is approximated by $(2\mathbb{1} - S)$, is

$$\tilde{\rho}(\mathbf{r}) = 2 \sum_{i,j=1}^{N} \chi_i(\mathbf{r})(2\delta_{ij} - S_{ij})\chi_j^*(\mathbf{r}). \tag{28}$$

This function satisfies $\tilde{\rho}(\mathbf{r}) = \rho(\mathbf{r})$ for the proper orthonormal solution, and is defined for all $\chi_i(\mathbf{r})$. As shown by Mauri *et al.*,[17] the Kohn-Sham energy Eq. (9) in which the band structure energy is computed using the functional $\tilde{E}$, and the charge density is approximated by Eq. (28), has a minimum for the orthonormal Kohn-Sham solution, and can therefore be used to find the self-consistent ground state. Some problems can appear, though, since the electron number is not conserved and one may find $\tilde{\rho}(\mathbf{r}) < 0$ for some S. This can only happen, however, in the initial steps of the minimization if a very bad initial guess is made. If a reasonable initial guess is used, no problems should appear. Furthermore, in a few iterations of the minimization of $\tilde{E}$, the wave functions become more orthonormal, and so the process is stable. At the end of the minimization the electron density will be the exact one, and the number of electrons will be the correct value, $2N$.

The present formulation is of general interest, not only for the linear scaling algorithms, but also for the standard iterative minimization methods with extended wave functions. The advantage of the new functional is that, being easy to compute, it avoids the need of the orthonormalization step. When used with extended wave functions (i.e., without imposing the localization condition), the method still scales as N^3. In order to fully exploit the present formulation and obtain an order-N scaling, we

turn now to use localized wave functions. As discussed in Sec. III, when the wave functions are localized, each of them only interacts with those in the same region, a number independent of the system size, leading to order-N scaling. This scaling is preserved regardless of the type of basis functions used to express the LWF's:

$$|\chi_i\rangle = \sum_{\mu=1}^{M} C_{\mu i}|\phi_\mu\rangle. \tag{29}$$

If we use localized basis orbitals (atomic orbitals or Gaussian functions) the finite localization of each LWF is easily achieved by expanding it only in terms of those $|\phi_\mu\rangle$ which are inside the localization radius R_c, so that there is only a limited number (independent of the size of the system) of $C_{\mu i} \neq 0$ for each LWF $|\chi_i\rangle$. The Hamiltonian and overlap matrix elements between LWF's

$$H_{ij} = \langle\chi_i|\hat{H}|\chi_j\rangle = \sum_{\mu,\nu=1}^{M} C^*_{\mu i} h_{\mu\nu} C_{\nu j} \ , \tag{30}$$

$$S_{ij} = \langle\chi_i|\chi_j\rangle = \sum_{\mu,\nu=1}^{M} C^*_{\mu i} s_{\mu\nu} C_{\nu j} \tag{31}$$

are given in terms of $h_{\mu\nu} = \langle\phi_\mu|\hat{H}|\phi_\nu\rangle$ and $s_{\mu\nu} = \langle\phi_\mu|\phi_\nu\rangle$, the matrix elements in the basis orbitals (which are zero for orbitals centered on distant sites). The number of nonzero matrix elements H_{ij} and S_{ij} is proportional to the system size, while the cost of the calculation of each of them is independent of the size of the system, since in the sums in Eqs. (30) and (31) one considers only those basis orbitals $|\phi_\mu\rangle$ and $|\phi_\nu\rangle$ within the localization range of $|\chi_i\rangle$ and $|\chi_j\rangle$, respectively, and that interact with each other via $h_{\mu\nu}$ or $s_{\mu\nu}$. Therefore the calculation of all the nonzero matrix elements is an order-N operation. Obviously, the computation of $\tilde{E}$ in Eq. (21) is also order-N.

If plane waves are used, the localization of the wave functions can be achieved with the methods proposed by Galli and Parrinello[13] (which perform the calculation dealing with the inversion of the overlap matrix that we avoid). In that case, the Hamiltonian and overlap matrix elements and the charge density can be calculated in a small mesh around each LWF, which using fast Fourier transform algorithms would require an effort proportional to N.

V. PRACTICAL IMPLEMENTATION OF THE ORDER-N SCHEME WITH A LOCALIZED BASIS

As we described in Sec. IV, when localized basis orbitals are used, an order-N scaling obtains. In the calculations, both the CPU time and the memory storage are optimized using sparse matrix multiplication routines, since only the nonzero elements of the product matrix are calculated and stored.

The search for the minimum of the energy functional as a function of the expansion coefficients $C_{\mu i}$ can be performed with several methods,[2] such as Car-Parrinello-type molecular dynamics or conjugate gradients. In both cases one needs to calculate the gradients of the energy functional at a given point. From Eqs. (21), (30), and (31), the gradients can be expressed as

$$\frac{\partial\tilde{E}}{\partial C^*_{\mu i}} = 4\sum_{\nu=1}^{M} H_{\mu\nu}C_{\nu i} - 2\sum_{\nu=1}^{M}\sum_{j=1}^{N} S_{\mu\nu}C_{\nu j}H_{ji} - 2\sum_{\nu=1}^{M}\sum_{j=1}^{N} H_{\mu\nu}C_{\nu j}S_{ji}. \tag{32}$$

which, again, can be calculated performing sparse matrix multiplications. Note, however, that not all the nonzero elements $\partial\tilde{E}/\partial C^*_{\mu i}$ have to be calculated, only those corresponding to the coefficients $C_{\mu i}$ that we allow to be nonzero (i.e., those within the localization radius for each LWF).

If the minimization is performed following the conjugate-gradients scheme,[2] for each step (n) of the minimization one must calculate the gradient at the current point $C^{(n)}_{\mu i}$ using Eq. (32), and then determine the minimization direction $D^{(n)}_{\mu i}$, which is conjugate to all the former directions. The energy has then to be minimized along the line defined by $C_{\mu i} = C^{(n)}_{\mu i} + \lambda D^{(n)}_{\mu i}$. It is very convenient to notice that the energy functional $\tilde{E}$ is a quartic function of the coefficients $C_{\mu i}$, if the Hamiltonian is not self-consistent, or if it is held fixed during the line minimization. Therefore, the *exact* line minimum can be found, as proposed in Ref. 15, by computing the energy at five different points of the line. One can also use the values of the energy and the gradient at the initial point ($\lambda = 0$) which are already calculated, and compute only the energy for three other values of λ. This procedure has the advantage that it avoids the need for choosing arbitrary "time steps," and assures the convergence towards the minimum.

In order to show the linear scaling of the algorithm, we have performed a series of calculations for silicon supercells in the diamond structure with different numbers of atoms, up to 1000. We have used an orthogonal tight-binding model with first-neighbor interactions.[42] Only the Γ point was used in the calculations. The LWF's were centered on bonds. In Fig. 4 we show the scaling of the CPU time with the number of atoms in the supercell for two different values of R_c, corresponding to LWF's confined to $n = 26$ and $n = 38$ atoms, respectively. The number of conjugate-gradients iterations was restricted to ten in all cases. As we will discuss in Sec. VIII, this is a typical number of iterations in molecular-dynamics simulations. We observe that the CPU time scales linearly with the number of atoms in the supercell. Obviously, the time increases with the number of atoms included in each localized function. For comparison, we also include the CPU time for exact diagonalization, which clearly shows a superlinear scaling.

Once the ground-state electronic energy has been calculated by means of conjugate gradients or other tech-

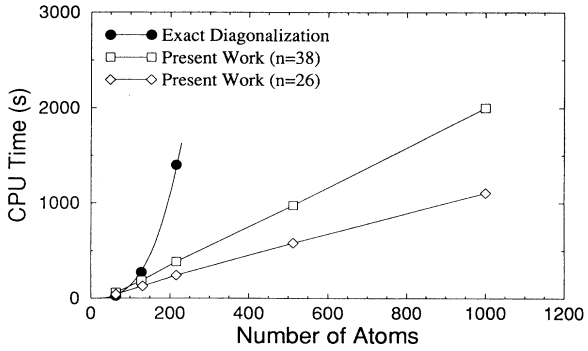

FIG. 4. CPU time vs number of atoms in the supercell for c-Si. We show the results for exact diagonalization and for our order-N scheme with localized functions confined to $n = 26$ and $n = 38$ atoms.

nique, one can find the forces between the atoms. These are needed for performing molecular-dynamics simulations, or looking for the equilibrium atomic configurations. We will assume that we know the variation of the Hamiltonian and the overlap matrix elements in the basis of atomic orbitals, so that we can calculate $\partial h_{\mu\nu}/\partial \mathbf{R}_l$ and $\partial s_{\mu\nu}/\partial \mathbf{R}_l$. This is clearly the case for the empirical tight-binding method, and for the Harris functional non-self-consistent version of the LDA that we will use in Sec. VII. As suggested by Sankey and Niklewski,[41] the band structure force can be evaluated using a variation of the Hellmann-Feynman theorem. Let us first derive the forces for the traditional energy functional Eq. (7). In that case,

$$\mathbf{F}_\mathbf{l}^{\mathrm{BS}} = -\frac{\partial \mathbf{E}_{\mathrm{BS}}}{\partial \mathbf{R}_\mathbf{l}} = -\mathbf{2}\mathrm{Tr}\left[\frac{\partial \mathbf{S^{-1}}}{\partial \mathbf{R}_\mathbf{l}}\mathbf{H} + \mathbf{S^{-1}}\frac{\partial \mathbf{H}}{\partial \mathbf{R}_\mathbf{l}}\right]. \quad (33)$$

To calculate $\partial S^{-1}/\partial \mathbf{R}_l$ we can take into account that $S^{-1}S = \mathbb{1}$, so that

$$\frac{\partial S^{-1}}{\partial \mathbf{R}_l} = -S^{-1}\frac{\partial S}{\partial \mathbf{R}_l}S^{-1}. \quad (34)$$

From Eqs. (30) and (31), we obtain

$$\frac{\partial H}{\partial \mathbf{R}_l} = C^\dagger \frac{\partial h}{\partial \mathbf{R}_l}C + \frac{\partial C^\dagger}{\partial \mathbf{R}_l}hC + C^\dagger h\frac{\partial C}{\partial \mathbf{R}_l}, \quad (35)$$

$$\frac{\partial S}{\partial \mathbf{R}_l} = C^\dagger \frac{\partial s}{\partial \mathbf{R}_l}C + \frac{\partial C^\dagger}{\partial \mathbf{R}_l}sC + C^\dagger s\frac{\partial C}{\partial \mathbf{R}_l}. \quad (36)$$

If we take into account that the band structure energy has been minimized with respect to the coefficients $C_{\mu i}$ so that $\partial E_{\mathrm{BS}}/\partial C^*_{\mu i} = 0$, after a cumbersome manipulation one finds that the terms involving $\partial C^\dagger/\partial \mathbf{R}_l$ and $\partial C/\partial \mathbf{R}_l$ from Eq. (35) cancel with those from Eq. (36), so that the forces read

$$\mathbf{F}_\mathbf{l}^{\mathrm{BS}} = -2\mathrm{Tr}\left[S^{-1}C^\dagger\frac{\partial h}{\partial \mathbf{R}_l}C\right] + 2\mathrm{Tr}\left[S^{-1}C^\dagger\frac{\partial s}{\partial \mathbf{R}_l}CS^{-1}H\right]. \quad (37)$$

The derivation of the forces for the family of functionals Eq. (26) follows the same lines:

$$\mathbf{F}_\mathbf{l}^{\mathrm{BS}} = -\frac{\partial \mathbf{E}^{(\mathcal{N})}}{\partial \mathbf{R}_\mathbf{l}} = -2\mathrm{Tr}\left[\frac{\partial\left(\sum_{m=0}^{\mathcal{N}}(\mathbb{1}-S)^m\right)}{\partial \mathbf{R}_l}H + \sum_{m=0}^{\mathcal{N}}(\mathbb{1}-S)^m\frac{\partial H}{\partial \mathbf{R}_l}\right]. \quad (38)$$

In this case, the derivative in the first term is

$$\frac{\partial}{\partial \mathbf{R}_l}\left(\sum_{m=0}^{\mathcal{N}}(\mathbb{1}-S)^m\right) = \sum_{m=1}^{\mathcal{N}}\sum_{n=0}^{m-1}(\mathbb{1}-S)^n\frac{\partial S}{\partial \mathbf{R}_\mathbf{l}}(\mathbb{1}-S)^{m-1-n}. \quad (39)$$

Again, expressing H and S in terms of the matrices C, $C^\dagger$, h, and s, and taking into account that $\partial E^{(\mathcal{N})}/\partial C^*_{\mu i} = 0$, one obtains

$$\mathbf{F}_\mathbf{l}^{\mathrm{BS}} = -2\mathrm{Tr}\left[\sum_{m=0}^{\mathcal{N}}(\mathbb{1}-S)^m C^\dagger\frac{\partial h}{\partial \mathbf{R}_l}C\right] + 2\mathrm{Tr}\left[\sum_{m=1}^{\mathcal{N}}\sum_{n=0}^{m-1}(\mathbb{1}-S)^n C^\dagger\frac{\partial s}{\partial \mathbf{R}_l} \times C(\mathbb{1}-S)^{m-1-n}H\right]. \quad (40)$$

It is easy to see that Eq. (40) reduces to Eq. (37) when $\mathcal{N} \to \infty$, since $S^{-1} = \sum_{m=0}^{\infty}(\mathbb{1}-S)^m$. We also recover the formulas derived by Sankey and Niklewski[41] for the orthonormal case $S = \mathbb{1}$. For the simplest functional $\mathcal{N} = 1$, which we will use in our calculations, the forces are

$$\mathbf{F}_l^{\mathrm{BS}} = -2\mathrm{Tr}\left[[\mathbb{1}+(\mathbb{1}-S)]C^\dagger\frac{\partial h}{\partial \mathbf{R}_l}C\right] + 2\mathrm{Tr}\left[C^\dagger\frac{\partial s}{\partial \mathbf{R}_l}CH\right]. \quad (41)$$

We see that the calculation of the forces reduces again to multiplications of sparse matrices, and can therefore be performed in order-N operations. Note that, as pointed out by Sankey and Niklewski,[41] Pulay corrections are not necessary, since we have taken derivatives of the matrix elements, not matrix elements of the derivatives.

VI. CONSEQUENCES OF THE LOCALIZATION

The localization of the wave functions can be viewed as a constraint that we impose on our minimization. The introduction of these "localization constraints" has several important consequences in the form of the solutions

and the properties of the energy functional.

(i) The energy minimum will be different from the exact ground-state energy, since a number of degrees of freedom has been removed (those $C_{\mu i}$ outside the localization range). This occurs because, as we discussed in Sec. III, the *exact* localized functions have long range tails that we are removing. We have, however, a variational approximation to the exact energy, since the minimum of the energy subject to the localization constraints is obviously higher than the unconstrained minimum. If the localization range is very large, this approximation will be very accurate, since the variables $C_{\mu i}$ that we will be neglecting would be in fact very small.

(ii) The solution that minimizes the energy $\tilde{E}$ (21) subject to the localization constraints will not be a strictly orthonormal set of states,[43] since the global minimum of $\tilde{E}$ cannot be reached in the constrained minimization. In particular, nonorthonormality is favored by the fact that nonorthonormal wave functions can be more localized than orthonormal solutions.[44] On the other hand, deviation from orthonormality is penalized by the second term of Eq. (21). A trade-off between these two effects is reached at the solution. The deviations from the orthogonality will be larger if the localization range is small, since in that case the energy gain will be more important.

(iii) The localization constraints cause the energy functional $\tilde{E}$ to have shallow local minima and flat regions in which the algorithm can be trapped for a long time during the minimization of the electronic wave functions. We have not found these problems if the wave functions are allowed to delocalize. In that case, it appears that the functional has a single minimum at the ground state of the system, or, more precisely, a locus of equivalent minima corresponding to unitary transformations of the ground state, as discussed in Appendix A. Imposing the localization constraints amounts to cutting the hypersurface defined by the functional $\tilde{E}$ in the space of parameters $C_{\mu i}$ with planes corresponding to the localization conditions $C_{\mu i} = 0$. It is clear that the intersection can have several local minima, even when the whole surface had just one. Moreover, the minima will now be disconnected, since we have eliminated part of the dimensions of the space. An example of these effects in a simple system is shown at the end of Appendix A.

(iv) Another consequence of the localization constraints clearly seen in the illustration in Appendix A is the symmetry breaking of the localized solutions. Although the *exact* Wannier-type wave functions can be made to have the symmetry of the system, when a localization cutoff is imposed the solution that minimizes the energy happens to break the symmetry. This point is further discussed in Appendix A. In realistic applications, like the Si and C systems discussed in Secs. VII and VIII, this situation also appears. If we are constructing localized functions centered in bonds, the lowest-energy solution consists of functions which are nonsymmetric with respect to the two atoms in the bond, each of them being slightly different from those centered on other bonds. To arrive at this symmetry-broken solution we must start the minimization with an initial guess that breaks the symmetry, otherwise we will obtain a symmetric solution with higher energy. To that end, we can add some small random numbers to the initial vectors, or give random small displacements to the atoms to slightly break the crystal symmetry. Clearly, this only applies to systems which have an internal symmetry, like crystals in the equilibrium configuration. For disordered systems, or crystals with thermal motion, the atomic disorder naturally breaks the symmetry of the wave functions, and no perturbation on the initial guess is necessary.

VII. ACCURACY OF THE SOLUTIONS VERSUS LOCALIZATION RADIUS

In this section we will study the accuracy of the solutions obtained with our order-N approach, as a function of the localization range imposed on the wave functions. First we will study the case of crystalline silicon, to test the accuracy of static properties such as the total energy, lattice constant, and bulk modulus. Then we will analyze the accuracy of the atomic forces in a disordered carbon system. In the applications of this section, we have used the local-orbital, *ab initio* total-energy molecular-dynamics method of Sankey and Niklewski.[41] The main approximations of this formulation are (1) four confined pseudoatomic orbitals (one s and three p's) per atom, (2) nonlocal, norm-conserving pseudopotentials,[45] and (3) the non-self-consistent Harris functional version of density-functional theory.[46] Note that using the Harris functional keeps the Hamiltonian fixed (no self-consistency is included), so, from a formal point of view, this is a very precise analog of the empirical tight-binding model. However, the basis is nonorthogonal and the orbital interactions are obtained from first principles.

The first system we study is crystalline silicon in the diamond structure. The atomic basis orbitals used for Si have a confinement radius of $5a_B$, producing third-neighbor interactions and overlaps (29 interacting atoms for each silicon). We have used a 512-atom supercell, restricting the sampling of the Brillouin zone (BZ) to the Γ point. In this case, the localized functions have been centered on bonds. In order to break the symmetry of the bond-centered wave functions (as discussed in Sec. VI), we first obtain the solution for a slightly disordered cell, and use this as an initial guess for the perfect crystalline cell. In Fig. 5 are shown the results for the cohesive energy, lattice constant, and bulk modulus versus the localization radius of the wave functions. We show the percent error with respect to the exact solution obtained by diagonalization. For a given cutoff radius R_c we include in each localized function the two atoms forming the bond, and all the atoms which are closer than R_c to either of these two atoms. The results show that the error in the cohesive energy decreases monotonically with increasing localization cutoff, as expected from the variational character of the energy functional $\tilde{E}$. The accuracy is excellent for R_c larger than 3 Å (i.e., including 26 or more atoms in each LWF), the error in the cohesive energy being smaller than 0.14 eV/atom. The lattice con-

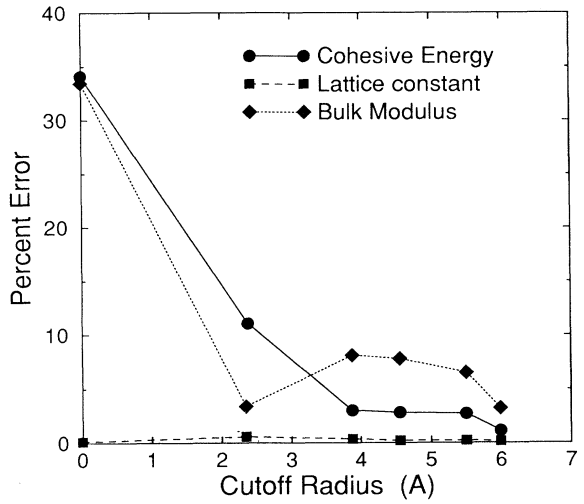

FIG. 5. Fractional errors vs localization cutoff R_c for a 512-atom supercell of c-Si. The LWF's were centered on bonds.

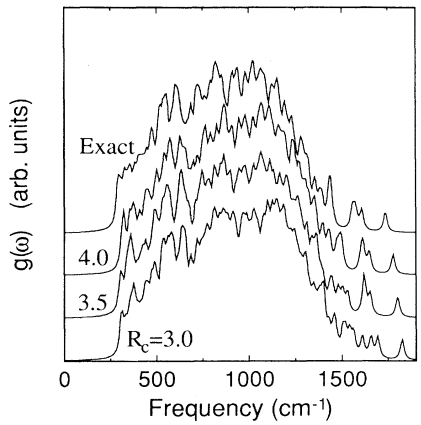

FIG. 6. Vibrational power spectrum for a 64-atom tetrahedral amorphous carbon cell. Each curve is obtained using a different value of the cutoff of the localized wave functions R_c (indicated by each curve), except the one labeled "Exact," which was obtained using exact diagonalization. The localized functions were centered on the atoms.

stant seems to be very insensitive to the cutoff R_c, giving errors of less than 1% in all cases. The bulk modulus is the most sensitive quantity as expected, since it involves the second derivative of the energy. However, the results are very acceptable, the error being lower than 10% for R_c larger than 2 Å.

In order to test the accuracy of the forces versus the localization cutoff of the electronic wave functions, we have performed a calculation of the vibrational spectrum of a 64-atom cell of tetrahedral amorphous carbon (*ta*-C). The cell was generated by Drabold *et al.*[47] The most important feature of this cell is that it models an amorphous tetrahedral network, and contains three types of structural defects: sp^2 atoms forming π bonds, strain defects with bonds angles larger than 150°, and stretched bonds longer than 1.8 Å. We therefore expect a large variety of localized vibrational modes together with the extended modes. The accuracy in the description of the localized modes will depend on the accuracy of the forces on the atoms in which each mode is localized. The atomic orbitals used for carbon have a confinement radius of $4.1a_B$. The Γ point was used to sample the BZ. The vibrational spectrum is computed by diagonalizing the dynamical matrix which is built computing the force acting on every atom in the system when each single atom is displaced a small distance from its equilibrium position.[48] Since only small displacements of one atom at a time are involved, the minimization of the energy is extremely efficient in this case, taking fewer than five iterations for each atomic displacement (using as the initial guess the solution for the undistorted system). In Fig. 6 we show the vibrational power spectrum for different localization ranges for the wave functions, together with the exact result. In this case, the wave functions were centered on atoms (two LWF's per atom). The results are in excellent agreement with the exact spectrum, being better for larger R_c. In all cases the spectral limits of the extended states band are accurately reproduced, as well as the peak structure in the high-frequency part of the spectrum, which is due to localized modes in the defects. A slight shift towards higher frequencies is observed for lower values of the localization radius, which indicates that the forces are larger than the exact ones when the localization range is small. The small differences in the spectra are probably due to the fact that the equilibrium structure of the cell is slightly different for different R_c. We therefore conclude that our order-N method produces forces which are accurate enough to describe a case as complex as the vibrational structure of an amorphous, defective network.

VIII. MOLECULAR DYNAMICS

In this section we apply our order-N approach to molecular-dynamics (MD) simulations. We will first discuss the general application of our energy functional to the standard quantum mechanical molecular-dynamics schemes, in the absence of the localization constraint. Then we will study the effects of the localization on the MD generated by our method. We analyze the accuracy of the dynamics that our scheme generates, and show that, even though some difficulties appear as a consequence of the localization constraints, the problems can be overcome by means of standard techniques.

A. MD algorithms

The application of our scheme to molecular-dynamics simulations is straightforward, given the simplicity of the formulas for the calculation of the atomic forces. There are essentially two different standard schemes for quantum mechanical MD simulations, both of which can be applied in the context of our order-N formulation.

(i) Car-Parrinello (CP) molecular dynamics.[1] In this method, the system of ions *and* electrons is governed by a

classical Lagrangian containing the physical ionic kinetic energy, the total electronic and ionic potential energies, and a fictitious electronic kinetic energy. The electronic degrees of freedom then obey the classical equations of motion dictated by the fictitious Lagrangian. The ionic and electronic "trajectories" are then integrated simultaneously with a finite difference method.[49] In this approach, the electrons are not in their ground state for each ionic position, but they can be made to stay close to the Born-Oppenheimer surface by choosing appropriately the initial conditions and the fictitious electronic mass. The advantage of the formulation is that the electronic energy need not be minimized for each ionic position, with the consequent savings in time. On the other hand, for the algorithm to be stable, the time step to integrate the equations of motion must be small.

In our order-N formulation, the expression for the CP Lagrangian takes the form

$$\mathcal{L} = \frac{1}{2}\mu \sum_{i=1}^{N}\sum_{\mu=1}^{M} \dot{C}_{\mu i}^2 + \frac{1}{2}\sum_{l=1}^{N_{\rm at}} M_l \dot{\mathbf{R}}_l^2 - E\left[\{C_{\mu i}, \mathbf{R}_l\}\right], \quad (42)$$

where the first term is the fictitious kinetic energy K_e of the electronic degrees of freedom, the second is the total ionic kinetic energy K_I, and the third is the electronic and ionic total potential energy, which includes the band structure energy functional Eq. (21). We have assumed that the electronic wave functions are real. μ is a fictitious electronic mass parameter, that is chosen to keep the electrons close to the ground-state surface. The classical Lagrange equations of motion determine the dynamics of the ions and the electronic variables. For the ions,

$$M_l \ddot{\mathbf{R}}_l = \mathbf{F}_l \ , \quad (43)$$

where $\vec{F}_l$ is the total force on ion l [which contains the band structure force Eq. (41)]. The electronic variables obey the equations of motion

$$\mu \ddot{C}_{\mu i} = -\frac{\partial E\left[\{C_{\mu i}, \mathbf{R}_l\}\right]}{\partial C_{\mu i}} \ . \quad (44)$$

The derivative of the energy with respect to the coefficients $C_{\mu i}$ is computed using Eq. (32). It is worth noticing that, in the present formalism, it is not necessary to include any orthonormality constraint term in the Lagrangian, since the energy functional takes care of the orthonormalization. In the standard CP techniques, one has to keep the functions orthonormal during the dynamics, either calculating the *exact* Lagrange multipliers for each time step, or performing an orthonormalization at the end of the step (see Ref. 50 for a detailed discussion), otherwise the total energy is not a constant of motion. In our case this is unnecessary, since our energy functional does not require orthonormal functions. During the simulation the wave function will depart slightly from orthonormality, but will oscillate about the orthonormal state, since it is the minimum of the energy functional Eq. (21). Since the equations of motion are derived directly from the Lagrangian, the internal total energy $U_{\mathcal{L}} = K_e + K_I + E[\{C_{\mu i}, \mathbf{R}_l\}]$ is a constant of motion for the trajectories generated by Eqs. (43) and (44). The total energy of the ionic system (the *physical* total energy $U_I = K_I + E[\{C_{\mu i}, \mathbf{R}_l\}]$) is not conserved, but, in systems with a gap, its value keeps very close to $U_{\mathcal{L}}$ and approximately constant, if the parameter μ is chosen small enough so that the characteristic frequency of the electronic degrees of freedom is much larger than the ionic frequencies.[51] In that case, the MD generated provide meaningful statistical averages in the microcanonical ensemble.

In order to illustrate the former scheme, we have performed a thermal simulation on a 64-atom cell of crystalline carbon in the diamond structure. We have used the tight-binding total-energy model of Xu *et al.*[52] using the Γ point to sample the Brillouin zone. We take a time step of 0.13 fs (5.37 atu), and a fictitious electronic mass $\mu = 300$ a.u. The ions are initially in the equilibrium crystalline positions, and the initial temperature is set to 400 K. The ionic trajectories were integrated using a fifth-order predictor-corrector algorithm,[49] and the electronic variables were solved with a velocity Verlet algorithm.[53] In Fig. 7 we show the total potential energy E, the physical total energy U_I, and the constant of motion $U_{\mathcal{L}}$, which is constant within the numerical accuracy. The physical total energy U_I is approximately constant, with oscillations of about 0.2 meV/atom, showing no significant drift during the simulation. The oscillations in the potential energies reflect the ionic thermal oscillations around the equilibrium crystal positions. These oscillations, as well as the ionic trajectories, are indistinguishable from those generated by a MD scheme with exact diagonalization, over periods of picoseconds. This shows that the energy functional $\tilde{E}$ is appropriate for molecular-dynamics simulations, and that, as we discussed above, it is not necessary to treat the orthonormalization of the wave functions during the simulation, in contrast with the standard iterative minimization methods.

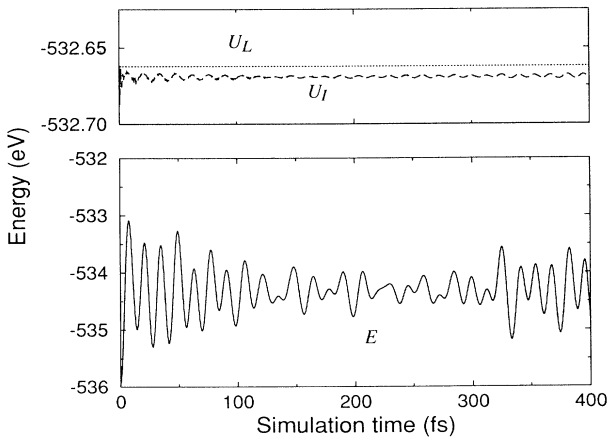

FIG. 7. Car-Parrinello molecular-dynamics simulation of diamond, with an initial temperature of 400 K. No localization constraints were imposed. We show the evolution of the total potential energy E (full line), total energy U_I (dashed line), and the Car-Parrinello constant of motion $U_{\mathcal{L}}$ (dotted line), as a function of time.

(ii) Conjugate-gradients dynamics. In this scheme, for a given atomic configuration, the electronic energy [Eq. (21) in our case] is minimized by means of a conjugate-gradients minimization. Once the electronic wave functions that minimize the electronic energy have been determined, the atomic forces [Eq. (41)] are computed, and the classical equations of motion for the ions are integrated by means of a Verlet algorithm or other finite difference method.[49] Within this approach the electrons are kept on the Born-Oppenheimer surface during the simulation. This yields accurate atomic forces and permits long time steps in the integration of the ionic equations of motion, typically one order of magnitude larger than those in CP dynamics. Under these circumstances, the total energy $U_I = K_I + E[\{C_{\mu i}, \mathbf{R}_l\}]$ is a constant of motion in the MD simulation. However, a strict tolerance must be imposed on the minimization of the electronic energy to obtain good accuracy, which can lead to a relatively large number of iterations during the conjugate-gradients minimization.

We have tested the CG MD scheme in the absence of localization constraints in the 64-atom crystalline carbon system discussed before. Again we use the Γ point to sample the BZ, and a fifth-order predictor-corrector algorithm to integrate the ionic equations of motion. Now, however, we use a larger time step of 1 fs (41.32 atu). At each time step of the simulation, the electronic energy $\tilde{E}$ was minimized within a tolerance of 10^{-8}. In order to reduce the number of iterations in each minimization, instead of using the solution of the former time step, we make a simple linear extrapolation[54] for the initial guess at time t_n, using the values of the wave functions at the times t_{n-1} and t_{n-2}:

$$C_{\mu i}(t_n) = C_{\mu i}(t_{n-1}) + [C_{\mu i}(t_{n-1}) - C_{\mu i}(t_{n-2})] \ . \quad (45)$$

This reduces the average number of iterations to about four. We show the results for the potential energy E and the total energy U_I in Fig. 8. We see that the total energy U_I is nearly constant throughout the simulation, the energy drift being smaller than 0.4 meV/atom ps. The evolution of the potential energy is identical to the one obtained in the Car-Parrinello simulation.

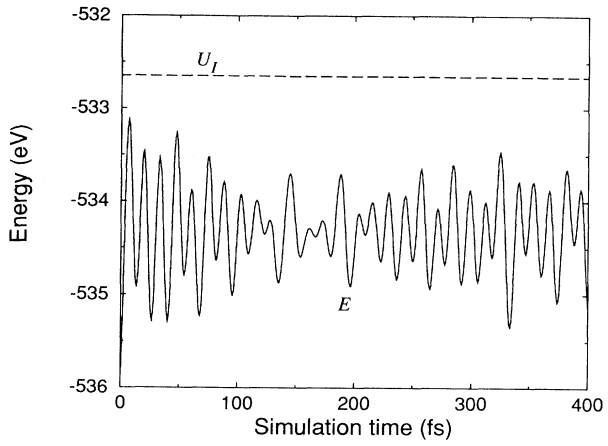

FIG. 8. Conjugate-gradients molecular-dynamics simulation of diamond, with an initial temperature of 400 K. No localization constraints were imposed. We show the evolution of the total potential energy E (full line) and the total energy U_I (dashed line) as a function of the time.

B. Molecular dynamics with localized wave functions

Once we have demonstrated the applicability of our energy functional to the standard MD schemes, we study the effect of the localization constraints on the dynamics generated by our order-N method. As we discussed in Sec. VI, when the electronic wave functions are required to be localized there are serious changes in the properties of the energy functional $\tilde{E}$ that will have an important effect in the molecular dynamics. In particular, the difficulty of minimizing $\tilde{E}$ is a serious obstacle in using the present scheme. If the number of iterations required to minimize the electronic energy to the desired tolerance is too large, we would be losing the advantages of the order-N scaling. This, however, can be avoided as is shown below. We have studied the same system as in Sec. VIII A, namely, the 64-atom supercell of carbon in the diamond structure with an initial temperature of 400 K. In all the cases, we have used wave functions centered on bonds, with a localization radius of 2.6 Å, for which there are 26 atoms in each LWF.

We start with the CG scheme. We have studied the quality of the MD generated using different tolerances in the minimization of the electronic energy $\tilde{E}$. In Fig. 9 we show the evolution of the total energy U_I with the simulation time for three different values of the relative tolerance $\tau = 10^{-8}$, 10^{-7}, and 10^{-6}. In all three cases, there is an appreciable drift in the total energy. For the most strict tolerance of $\tau = 10^{-8}$, we see that the drift is continuous, with an energy loss of about 19 meV/atom ps. In this case, the average number of CG iterations per simulation time step is 15, considerably larger than the one corresponding to the nonlocalized case for the same

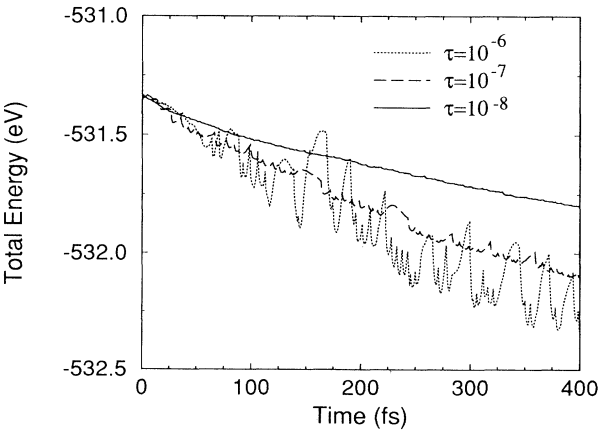

FIG. 9. Total energy U_I vs time for several conjugate-gradients MD simulations of diamond, with an initial temperature of 400 K. The localized functions are centered on bonds and are restricted to 26 atoms. We show the results for different tolerances in the CG minimization: 10^{-8} (full line), 10^{-7} (dashed line), and 10^{-6} (dotted line).

tolerance. For the less demanding tolerances of 10^{-7} and 10^{-6}, we see that the total energy suffers jumps with time and a larger drift. The jumps are due to the fact that the energy functional has local minima and flat regions, as discussed in previous sections. If the tolerance in the minimization is not strict enough, the electronic wave functions get trapped in a local minimum for some time, forcing the total energy to increase. Then, as the atoms move, the electronic variables are eventually able to escape the local minima into a lower-energy solution, and the total energy decreases suddenly. An analysis of the number of CG iterations per time step indicates that this number is maximum at the points of sudden decrease in the total energy (about five times larger than the average), showing that the wave functions are finding a path to another minimum.

One can, alternatively, use a fixed number of CG iterations per time step, instead of minimizing the electronic energy up to a certain tolerance. We show the results of such an approach in Fig. 10, for different numbers of CG iterations. We can see that, with 100 iterations per time step, there is still an energy loss, but it reduces to 4 meV/atom ps. The energy drift increases with decreasing number of iterations, as expected. However, we see that, even in the case of 10 iterations per step, the energy curve is smooth, in contrast to the case in which we imposed a certain tolerance in the minimization. This means that the wave functions are able to escape from local minima during the simulation and to relax to more stable regions. We have therefore found a way to overcome the difficulty of the local minima in the energy functional $\tilde{E}$. However, only when the number of iterations is large are the forces accurate enough to produce dynamics with a small energy loss. This would be a drawback in the application of the method to MD, since a large number of iterations would obviously decrease the efficiency greatly.

In order to understand the origin of this energy loss and cure this problem, we analyze the MD generated with the Car-Parrinello scheme, in the presence of localization constraints. In Fig. 11 we show the results for the total potential energy E, the physical total energy U_I, and the constant of motion $U_{\mathcal{L}}$. Whereas $U_{\mathcal{L}}$ is constant throughout the simulation, indicating that the integration of the equations of motion was accurate, there is a total energy U_I loss of 13 meV/atom ps, similar to the CG dynamics. The difference between $U_{\mathcal{L}}$ and U_I is the electronic fake kinetic energy. We notice that there is a continuous heating of the electronic system, at the expense of the decrease of energy in the ionic system (see Fig. 11). This means that there is a transfer of heat between the ionic and electronic systems, which makes the electronic wave functions depart from the Born-Oppenheimer surface. In the standard Car-Parrinello scheme, this situation is typical of systems in which the gap for the electronic excitations is very small or zero.[51] In this case, the characteristic frequencies of the electronic system overlap with those of the ions, and the adiabatic approximation breaks down, resulting in an energy transfer from the ions to the electronic variables. In our case, even if the system has a gap, the localization constraints introduce small frequency modes in the electronic energy

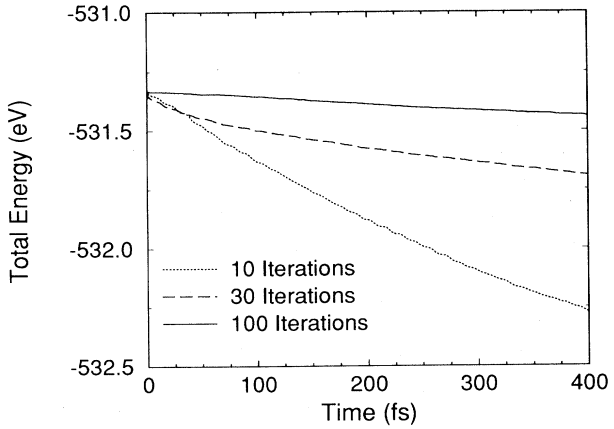

FIG. 10. Total energy U_I vs time for several conjugate-gradients MD simulations of diamond, with an initial temperature of 400 K. The localized functions are centered on bonds and are restricted to 26 atoms. We show the results for a fixed number of CG iterations per time step: 100 (full line), 30 (dashed line), and 10 (dotted line) iterations.

surface (flat energy regions) that are responsible for the energy transfer.[55] The similarity in the dynamics of our formulation with localization constraints and the standard Car-Parrinello scheme in metals suggests a possible way to overcome the energy loss of the ionic system to the electronic wave functions. As proposed by Blöchl and Parrinello,[56] one can avoid the energy transfer by coupling two Nosé thermostats[57–60] acting as heat baths that keep each system at a different temperature. The ionic thermostat is set to the physical temperature wanted for the simulation, whereas the electronic thermostat is kept to a small temperature, to avoid the heating of the electronic system. The ionic dynamics generated within this scheme are not microcanonical (since there is an energy flow between the system and the thermostats) but rather belong to the canonical ensemble. We have applied this

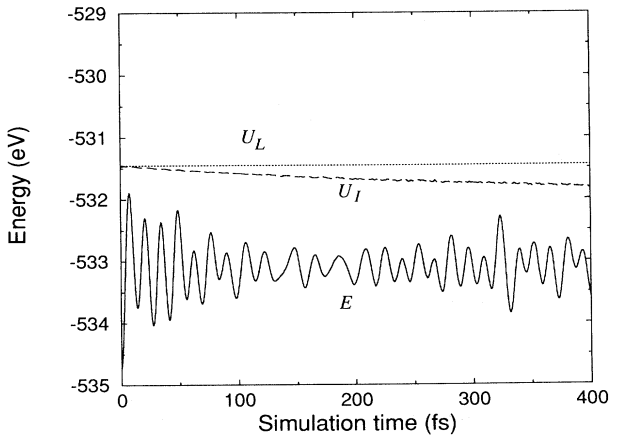

FIG. 11. Total potential energy E (full line), total energy U_I (dashed line), and the Car-Parrinello constant of motion $U_{\mathcal{L}}$ (dotted line) as a function of the time for a Car-Parrinello MD simulation of diamond, with an initial temperature of 400 K. The localized functions are centered on bonds and are restricted to 26 atoms.

procedure to our scheme, and show the results in Fig. 12. The temperature of the thermal bath for the ions has been set to 200 K (which is the approximate equilibration temperature that the system reached in the previous examples), and the target electronic kinetic energy was 8 meV. We see that now the physical total energy U_I is not conserved, but it oscillates around the mean value without any appreciable drift. The oscillations are due to fluctuations in the energy in the canonical ensemble. On the other hand, the electronic kinetic energy does not rise, since the thermal bath absorbs the excess of energy transferred from the ionic system.

We can use the same kind of ideas in the CG molecular dynamics. Here, we do not need to use any Nosé thermostat for the electrons, since they are minimized to (or close to) the Born-Oppenheimer surface during the simulation. However, in order to avoid the energy loss in the ionic system due to the lack of accuracy of the atomic forces (especially when a small number of iterations is used), we can couple a Nosé thermostat to the ions, set to the physical temperature wanted in the simulation. This is done in Fig. 13 for two different values of the number of CG iterations per time step. We see that in both cases the total energy oscillates around the average value with no overall drift, as in the case of Car-Parrinello MD shown in Fig. 12. Notice that the dynamics generated with 10 iterations per time step are essentially identical to those generated with 30 iterations, showing that, once the total-energy drift has been controlled with the Nosé thermostat, the CG MD can be performed using a small number of iterations. This is of paramount importance for the practical application of the order-N scheme to realistic MD simulations. We also note that the results are very similar to those obtained with Car-Parrinello dynamics (with two Nosé thermostats) shown in Fig. 12, indicating

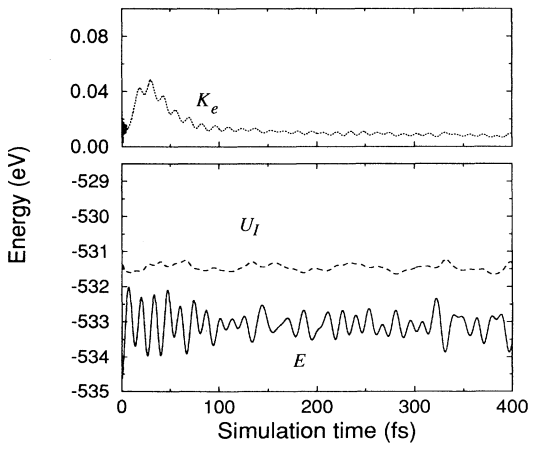

FIG. 12. Total potential energy E (full line), total energy U_I (dashed line), and fake electronic kinetic energy K_e (dotted line) as a function of the time for a Car-Parrinello MD simulation of diamond with two Nosé thermostats (one for the electrons and one for the ions). The initial temperature was 400 K; the target ionic temperature was 200 K and the target electronic kinetic energy 8 meV. The localized functions are centered on bonds and are restricted to 26 atoms.

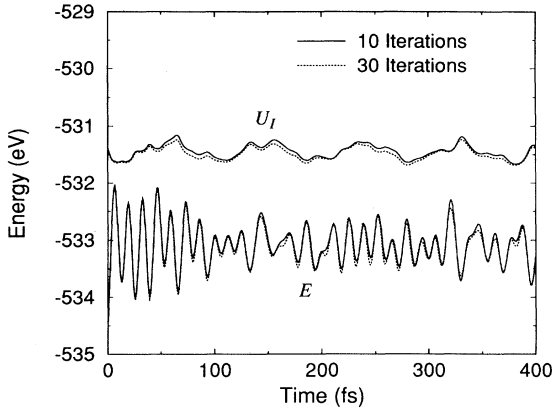

FIG. 13. Conjugate-gradients MD simulation of diamond with a Nosé thermostat coupled to the ionic system. The initial temperature was 400 K and the target ionic temperature was 200 K. The localized functions are centered on bonds and are restricted to 26 atoms. We show the total potential energy E and the total energy U_I as a function of time for two simulations with 10 and 30 CG iterations per time step, respectively.

that the results are robust and do not depend on the details of the simulation algorithm. From a practical point of view, the choice between Car-Parrinello and conjugate-gradients dynamics depends mostly on the computational difficulty of calculating the atomic forces, depending on the particular model being used. Since the CP approach requires the calculation of the forces more often (because the time step is smaller), this method would be preferred when this computation is cheap, as in empirical tight-binding models. However, if the forces involve a large numerical effort, the CG scheme is more economical.

The use of the Nosé thermostat allows us to perform reliable order-N MD simulations, either within the Car-Parrinello scheme or with conjugate gradients with a small number of iterations. The dynamics generated correspond to the canonical ensemble, not the microcanonical. This has advantages if we want to perform a thermal annealing, or a thermal simulation, for example, since in that case the calculation resembles the physical situation. However, in certain cases, one may be interested in working within the microcanonical ensemble, i.e., at constant total energy. This is the case, for example, for the simulation of a phonon mode, in which the system is not at thermal equilibrium. In this case, using a thermal bath would not be appropriate. Here we propose a way to perform constant-energy MD simulations, avoiding the energy loss. The procedure is inspired by the success of the Nosé thermostat in producing reliable CG dynamics even when the forces are not accurate due to the small number of iterations. In particular, the fact that, using the Nosé thermostat, the dynamics are essentially identical when the number of iterations is 10 (where the atomic forces have larger errors) or 30 iterations suggests that the errors in the forces are random and that on supplying the energy lost (with the Nosé thermostat) the difference in the trajectories is negligible. In this case, we could perform constant-energy MD simulations, restoring the

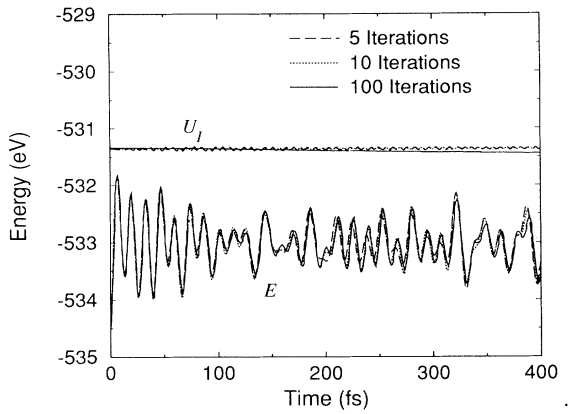

FIG. 14. Conjugate-gradients MD simulation of diamond. We show the total potential energy E and the total energy U_I for two simulations with five (dashed line) and ten (dotted line) iterations per time step, using an "energy thermostat" to conserve the total energy (see text). For comparison, we also show the results using 100 iterations (full line) without any thermostat. In all cases the initial temperature was 400 K and the localized functions are centered on bonds and are restricted to 26 atoms.

energy lost in each time step by simply rescaling the velocities in such a way that the total energy U_I is constant. However, this could introduce discontinuous perturbations on the atomic movements, so we have used a more refined procedure. We use the same formalism as in the constant-temperature Nosé dynamics, but with a target ionic temperature that varies with time. The target ionic kinetic energy at each time step is chosen as the kinetic energy that would make the total energy constant. The effect of this "energy thermostat" is to restore the missing energy to the ionic system, if some energy is lost during the simulation, but in a smooth way. On the other hand, if the system has no energy losses (i.e., the forces are accurate), then the instantaneous kinetic energy at each time is equal to the target kinetic energy (since the energy is conserved), and the forces on the atoms due to the "thermostat" would be zero throughout the simulation. We have tested this approach in the diamond system discussed above. In Fig. 14 we show the results using five and ten CG iterations per simulation step, and compare them with the case in which the number of iterations was 100 and no thermostat was used. We see that the "energy thermostat" is effective in keeping the total energy close to a constant value. The small oscillations are due to the dynamic nature of the Nosé algorithm. It is important to notice that the potential energy shows the same behavior in all three cases, showing that the use of the "energy thermostat" allows us to obtain, with a small number of iterations, the same dynamics as we obtain with as many as 100 iterations per time step. An analysis of the trajectories leads to the same conclusion. In any case, the errors involved in using the thermostat and a small number of iterations are much smaller than the typical errors due to the truncation of the localized wave functions, as can be seen by comparing with the curves in Figs. 7 and 8.

IX. CONCLUSION

We have introduced and rigorously analyzed a practical scheme for computing total energies and forces from the electronic structure with linear system-size scaling. The essential approximation of the method is the use of the electronic truncation radius R_c, which is completely controllable, and we recover exact results in the large-R_c limit. An energy functional was introduced that allows the calculation of the ground-state energy without the necessity of explicitly orthogonalizing the wave functions. We tested the method on diamond and tetrahedral amorphous carbon and found close agreement with conventional calculations. We have found that the most delicate part of the calculation is connected with truncating the Wannier-like representation of the electronic states, and described the error accruing from this truncation. Two schemes were discussed for performing dynamical simulation, based on conjugate-gradient and Car-Parrinello dynamics. Due to the difficulty in minimizing the energy functional, and the consequent errors in the ionic forces, we found that a Nosé thermostat produces realistic simulations without staying exactly on the Born-Oppenheimer surface. Reliable molecular dynamics can be performed in this way with a small number of iterations in the electronic energy minimization, therefore preserving the advantages of the order-N scaling.

The present approach has features which can be useful in other contexts. For example, the unconstrained minimization could be useful in any of the current methods[2] to avoid explicit orthogonalization. Similarly, the method to reduce the effects of the errors in the forces can be used in current MD methods which rely upon accurate electronic minimization at each step.[2] Finally, the construction of Wannier (or Wannier-like) functions is useful in itself.

ACKNOWLEDGMENTS

We would like to acknowledge informative discussions with D. Vanderbilt, O. Sankey, G. Galli, F. Mauri, and E. Stechel. We also wish to thank Dmitry Lebedenko and Satoshi Itoh for their help in some of the stages of this work. One of us (P.O.) acknowledges support from DGICyT (Spain). This work was partially supported by the National Science Foundation under Grants No. DMR-89-20538, No. DMR-93-22412, and No. DMR-91-21570.

APPENDIX A: PROPERTIES OF THE ENERGY FUNCTIONAL

In this appendix we will study the stability of the new energy functional $\tilde{E}$. We will describe under what conditions the energy functional $\tilde{E}$ has a lowest bound (i.e., a global minimum) or at least a local minimum at the correct position (i.e., the subspace of the occupied states). We will also illustrate the features of the localized solu-

tions, discussed in Sec. VI, with a simple example.

In this appendix we will work in terms of the eigenvectors of the Hamiltonian $\hat{H}$ (which, for the sake of simplicity, will be non self-consistent). If the Hamiltonian is defined in an M-dimensional space (where M will be the size of the basis set), we will define the set $\{|\psi_i\rangle, i = 1, ..., M\}$ as the (orthonormalized) eigenvectors of $\hat{H}$, with eigenvalues ϵ_i, which will be assumed to be set in growing order. In that case, the correct band structure energy equals the trace of $\hat{H}$ in the occupied subspace, i.e., the sum of the N lowest eigenvalues of $\hat{H}$.

As described in Sec. IV, our energy functional is defined, for an arbitrary set of N nonorthonormal states $\{|\chi_i\rangle, i = 1, ..., N\}$, by Eq. (21). To study the properties of our functional, it is useful to express the states $|\chi_i\rangle$ in terms of the eigenvectors of $\hat{H}$ (this can be done since the eigenvectors are a basis of the space in which $\hat{H}$ is defined):

$$|\chi_i\rangle = \sum_{j=1}^{M} A_{ij}|\psi_j\rangle, \qquad i = 1, ..., N. \tag{A1}$$

Using this expression in Eq. (21) and taking into account that $|\psi_i\rangle$ are orthonormal eigenstates of $\hat{H}$, we can express the energy functional $\tilde{E}$ as

$$\tilde{E} = 2\sum_{j=1}^{M} \epsilon_j - 2\sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \left[\sum_{i=1}^{N} A_{ij}A_{ik} - \delta_{jk}\right]^2 . \tag{A2}$$

From this expression we can already extract some of the properties of the energy functional $\tilde{E}$

(i) $\tilde{E}$ takes the correct value at the ground state of the system; clearly, if the set of states for which $\tilde{E}$ is evaluated corresponds to the lowest eigenvectors of $\hat{H}$, then

$$A_{ij} = \begin{cases} \delta_{ij}, & j = 1, ..., N, \\ 0, & j = N+1, ..., M , \end{cases} \tag{A3}$$

and the value of our functional is the exact band structure energy, Eq. (7):

$$\tilde{E} = 2\sum_{i=1}^{N} \epsilon_i. \tag{A4}$$

(ii) The functional $\tilde{E}$ is invariant under unitary transformations. If we apply a unitary transformation B on the vectors $|\chi_i\rangle$ to obtain another set of vectors

$$|\chi_i'\rangle = \sum_{j=1}^{N} B_{ij}|\chi_j\rangle, \qquad i = 1, ..., N, \tag{A5}$$

it is straightforward to see that the value of $\tilde{E}$ evaluated for the new set $\{|\chi_i'\rangle\}$ equals that for the set $\{|\chi_i\rangle\}$ (one has to use the fact that B is unitary, so that $BB^\dagger = \mathbb{1}$).

(iii) The energy functional $\tilde{E}$ has a lower bound if and only if all the eigenvalues of $\hat{H}$ are negative (i.e., the Hamiltonian is negative definite). If any of the eigenvalues are positive (suppose it is the highest level ϵ_M), $\tilde{E}$ has no lower bound, since we can build a set of vectors $|\chi_i\rangle$ of the form

$$\begin{aligned} |\chi_i\rangle &= |\psi_i\rangle \qquad (i = 1, ..., N-1) , \\ |\chi_N\rangle &= \eta|\psi_M\rangle \end{aligned} \tag{A6}$$

(which corresponds to an excited state of the system, where the $N-1$ lowest levels are occupied, and the Nth electron has been promoted to the state with positive eigenvalue, allowing this state $|\psi_M\rangle$ to have an arbitrary norm η) for which the energy functional diverges to $-\infty$ as $\eta \to \infty$:

$$\tilde{E} = 2\sum_{i=1}^{N-1} \epsilon_i - (\eta^4 - 2\eta^2)\epsilon_M \to -\infty . \tag{A7}$$

If, on the contrary, all the eigenvalues are negative, the functional $\tilde{E}$ clearly has a lower bound: the first term in Eq. (A2) is constant, and the contribution from the second term is always positive.

(iv) Since A is a rectangular matrix of dimensions $N \times M$, not all the terms inside the brackets of Eq. (A2) can be zero simultaneously, since it would mean that $A^\dagger A = \mathbb{1}$ (where $\mathbb{1}$ is the $M \times M$ unit matrix), which is not possible. One can see that this term reaches its minimum value when

$$\sum_{i=1}^{N} A_{ij}A_{ik} = \delta_{jk}, \quad j, k = 1, \ldots, N$$

and (A8)

$$A_{ij} = 0, \quad j = N+1, \ldots, M ,$$

in which case the states $|\chi_i\rangle$ are just any unitary transformation of the N lowest eigenvectors of $\hat{H}$ and therefore constitute an *orthonormal basis of the occupied subspace.* The important result that we have sketched is that, if all the eigenvalues of $\hat{H}$ are negative, $\tilde{E}$ has a *global* minimum; this global minimum, for which the value of $\tilde{E}$ is the ground-state band structure energy, is multiple and is located at any *orthonormal* set of states that span the occupied subspace of $\hat{H}$ (i.e., at the point defined by the N lowest eigenstates of $\hat{H}$ and at any unitary transformation of this set).

(v) Even if some of the empty eigenvalues of $\hat{H}$ are positive, the functional $\tilde{E}$ has a *local* minimum for the ground state of the system, provided that all the occupied eigenvalues are negative. To prove this we can calculate the value of the functional for a point in the neighborhood of the ground state. Such a point can be expressed, in terms of Eq. (A1), as

$$A_{ij} = \begin{cases} \delta_{ij} + \varepsilon_{ij}, & j = 1, ..., N \\ \varepsilon_{ij}, & j = N+1, ..., M , \end{cases} \tag{A9}$$

where ε_{ij} are small. The difference in the value of the

functional $\tilde{E}$ at this point and the ground state is, to second order in ε_{ij},

$$\tilde{E} - E_{\rm BS} = -2\sum_{i=1}^{N}\epsilon_i\sum_{j=1}^{N}(\varepsilon_{ij}+\varepsilon_{ji})^2 - 2\sum_{i=1}^{N}\sum_{m=N+1}^{M}(\epsilon_i-\epsilon_m)\varepsilon_{im}^2 \ . \qquad \text{(A10)}$$

Each term in this equation is non-negative if (1) the energies of the occupied eigenstates are equal to or less than the energies of the empty eigenstates (which is always granted by definition) and (2) the occupied eigenvalues have negative energy.

At points where the second-order expansion (A10) vanishes, we must go to higher terms in ε_{ij}. This happens when

$$\varepsilon_{ij} = \begin{cases} -\varepsilon_{ji}, & j = 1,...,N \\ 0, & j = N+1,...,M \ , \end{cases} \qquad \text{(A11)}$$

in which case one can verify that

$$\tilde{E} - E_{\rm BS} = -2\sum_{i=1}^{N}\epsilon_i\sum_{j=1}^{N}\left[\sum_{k=1}^{N}\varepsilon_{ki}\varepsilon_{kj}\right]^2 . \qquad \text{(A12)}$$

Again, this is non-negative if the occupied eigenvalues are negative. Therefore, we have shown that the functional $\tilde{E}$ has a *local* minimum at the correct ground state of the system if all the occupied eigenvalues are negative. If some of the occupied states have positive eigenvalues, an energy shift $\epsilon \to \epsilon - \eta$ is necessary to make the functional stable. This can be achieved by making the transformation $h_{\mu\nu} \to h_{\mu\nu} - \eta s_{\mu\nu}$ on the Hamiltonian matrix elements between the basis orbitals.

Furthermore, as an important consequence of point (ii), if a set of states $\{|\chi_i\rangle, i = 1,...,N\}$ is a global (or local) minimum of the functional $\tilde{E}$, all the unitary transformations $\{|\chi_i'\rangle, i = 1,...,N\}$ are also global (local) minima of $\tilde{E}$. The demonstration for a global minimum is trivial, since the energy is the same for the two sets. For the case of a local minimum, one can see that any point in the proximity of the set $\{|\chi_i'\rangle\}$ can be expressed as a unitary transformation of some point in the proximity of the set $\{|\chi_i\rangle\}$, so that if the latter is a local minimum then the former is also a local minimum.

In order to illustrate the consequences of the localization of the wave functions on the solution, as discussed in Sec. VI, we will consider a ring with four orbitals, labeled from 1 to 4. Orbitals 1 and 3 have on-site energy $\epsilon_A = -5$ eV, and orbitals 2 and 4, have $\epsilon_B = -4$ eV, and the first-neighbor hopping is $V = -1$ eV. We consider the half filling case, i.e., four electrons. We have, therefore, two occupied wave functions, $|\psi_1\rangle$ and $|\psi_2\rangle$, shown schematically in Fig. 15(a). As we have discussed, any unitary transformation of these eigenstates has the same energy. Therefore there exists an infinite number of solutions, all with the same energy, since the number of unitary transformations that one can perform is infinite (corresponding in this case to all the continuous

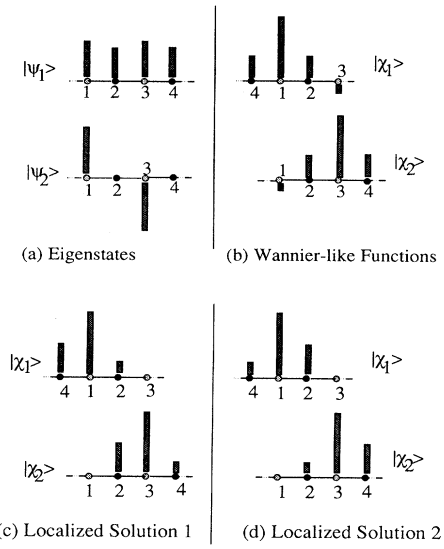

FIG. 15. Scheme of the wave-function solution of the four-atom ring. We show the weight of each wave function on the orbital of each atom for (a) eigenvectors; (b) Wannier functions; (c) and (d) two equivalent localized solutions (restricted to be localized to three atoms, and centered on atoms 1 and 3).

set of rotations of the two vectors). This is reflected in the fact that, if we look for the solution minimizing the energy functional Eq. (21), we find different solutions with the same energy depending on the initial vectors used in the minimization. Analytically, we can build the Wannier-type solution as the sum and difference of the two eigenvectors:

$$|\chi_1\rangle = \frac{1}{c}\left(|\psi_1\rangle + |\psi_2\rangle\right) \ , \qquad \text{(A13)}$$

$$|\chi_2\rangle = \frac{1}{c}\left(-|\psi_1\rangle + |\psi_2\rangle\right) , \qquad \text{(A14)}$$

where c is a normalization constant. These functions, shown in Fig. 15(b), are centered on atoms 1 and 3, respectively, and are symmetric. We now impose a localization range on the solution: we build one wave function centered on atom 1 and the other on atom 3, and we allow them to spread only to the neighbor atoms (i.e., the first function is zero on atom 3, and the second is zero on atom 1). We minimize the energy functional subject to the localization constraint, and find that there are *only two* equivalent global minima, shown in Figs. 15(c) and 15(d). Either of these two solutions is reached regardless of the initial choice of vectors used in the minimization, which means that these are the only global minima. The localized solutions clearly resemble the exact Wannier-type solutions shown in Fig. 15(b). Note that, when a localization cutoff is imposed, the solution that minimizes the energy breaks the symmetry, as we can see in Figs. 15(c) and 15(d). Note that, even though each of the localized wave functions does not obey the system

symmetry, nevertheless, the total charge density has the correct symmetry. For instance, atoms 1 and 3, which should be equivalent by symmetry, have the same charge, and the same is true for atoms 2 and 4.

APPENDIX B: CONNECTIONS TO DENSITY MATRIX METHODS

In this section we provide alternate derivations and discussions of the density matrix (DM) approach to electronic-structure calculations.[27,28] In this method for solving the electronic-structure problem, the orthogonality constraints are translated into idempotency conditions on the DM ρ. A functional is then introduced which possesses a minimum for idempotent ρ and minimum band energy. We discuss the properties of this functional and its connection to other work in this Appendix.

Calculations using the DM as a variationally determined object were done by McWeeny,[61] in which an always idempotent DM was varied to minimize an energy functional. McWeeny also introduced the concept of the "purification transformation:" a mapping which takes a nonidempotent matrix into a more idempotent form. The form of this map $f:\rho_i \rightarrow \rho_{i+1}$ is $\rho_{i+1} = 3\rho_i^2 - 2\rho_i^3$, which converges to an exactly idempotent matrix as $i \rightarrow \infty$. Effectively this map has a "fixed point" for idempotent ρ. McWeeny showed that this sequence is just the steepest descent minimization of the quantity $\mathrm{Tr}(\rho^2 - \rho)^2$. Smith and Gay[62] used the DM as a variationally determined object in solid state electronic-structure calculations on a Wannier function basis. They also introduced a successive approximation scheme for computing ρ.

Li, Nunes, and Vanderbilt,[27] and, independently, Daw[28] have shown how to exploit the McWeeny purification transformation to produce a linear scaling algorithm for electronic computations in solids. In their method they assume an orthonormal basis set, and propose an energy functional $E = \mathrm{Tr}[(3\rho^2 - 2\rho^3)(h - \mu)]$, for chemical potential μ. We call this the LNVD functional in what follows. Note that the trace of the LNVD functional is taken in the whole Hilbert space, i.e., the space spanned by the basis set, and not only the occupied space, as in our functional Eq. (21).

First, we provide a new derivation of the LNVD density matrix functional, and show how to convert the nonholonomic idempotency constraint into holonomic form, to obtain the form of the DM functional from a variational calculation. We need a functional which (1) minimizes the band energy $\mathrm{Tr}(\rho h)$, (2) has this minimum when ρ is idempotent, and (3) occupies exactly N states. Condition (2) is easily written as a holonomic constraint if we note that, if the basis is orthonormal, it is equivalent to $\rho_{\mu\nu} = \sum_\lambda \rho_{\mu\lambda}\rho_{\lambda\nu}$. We therefore apply the usual Lagrange method which is ideally suited for a holonomic constraint optimization problem, and therefore define the auxiliary functional

$$\tilde{E} = 2\Bigg[\mathrm{Tr}(\rho h) - \mu[\mathrm{Tr}(\rho) - N] - \sum_{\lambda\nu}\Lambda_{\lambda\nu}\Big[\rho_{\lambda\nu} - \sum_\sigma \rho_{\lambda\sigma}\rho_{\sigma\nu}\Big]\Bigg] \tag{B1}$$

for $\Lambda_{\mu\nu}$ Lagrange multipliers, with μ the chemical potential and N the number of electrons. Unconstrained variation of ρ, such that $\delta E/\delta\rho_{\mu\nu} = \delta E/\delta\Lambda_{\mu\nu} = 0$, implies that the matrix of Lagrange multipliers satisfies

$$h + \Lambda - \Lambda\rho - \rho\Lambda = 0\ . \tag{B2}$$

Since, at the solution, $\hat{\rho}$ is idempotent and commutes with $\hat{H}$, a solution to Eq. (B2) is $\Lambda = -(h - \mu)(2\rho - \mathbb{1})$. While this is only rigorously exact for the correct density matrix, it can be used even for ρ deviating from the correct solution, in a fashion similar to the discussion in Sec. IV for our orbital formulation. Substituting this Λ into Eq. (B1) yields the LNVD DM functional.

As currently implemented, the DM method requires initial guesses for the chemical potential μ and the DM. The former requires some global information about the density of states which is readily obtained by using recently proposed statistical methods.[23] This method yields an efficient order-N guess for the Fermi level. Similar methods can be used to estimate the full density matrix when used with approximate (polynomial) representations of the density operator.[24]

Next we prove that this DM functional possesses a local minimum at the correct ground state. Let ρ be the correct density matrix and δ be a small, arbitrary matrix. We study the energy of the proposed DM functional for small variations δ about the exact ρ and determine conditions for which E has a local minimum. We compute

$$\Delta\tilde{E} = \tilde{E}(\rho + \delta) - \tilde{E}(\rho). \tag{B3}$$

Simple manipulations show that to second order in δ we have

$$\Delta\tilde{E} = 4\ \mathrm{Tr}[\delta^2(\mathbb{1} - 2\rho)(h - \mu)] + 2\mathrm{Tr}[\delta(\mathbb{1} - 2\rho)\delta(h - \mu)]. \tag{B4}$$

There exists a unitary transformation U which simultaneously diagonalizes h and ρ. By trace invariance, we may evaluate $\Delta\tilde{E}$ in this diagonal representation to obtain

$$\Delta\tilde{E} = 4\sum_{\alpha\nu}\tilde{\delta}^2_{\alpha\nu}(1 - 2n_\alpha)(\epsilon_\alpha - \mu) + 2\sum_{\alpha\nu}\tilde{\delta}^2_{\alpha\nu}(1 - 2n_\nu)(\epsilon_\alpha - \mu), \tag{B5}$$

where ϵ_α are eigenvalues of h and n_α are the eigenvalues of the DM (the occupation numbers), and $\tilde{\delta} = U^T\delta U$. The DM calculations are implemented at fixed μ, which implies *some* N. For any such case, there exists a correct density matrix ρ about which we study the effect of the arbitrary perturbation. For variations about this ρ, we clearly have $\Delta\tilde{E} > 0$, since $\tilde{\delta}^2_{\mu\nu} > 0$ always, and for $\epsilon_\nu > \mu$ the occupation numbers n_ν are zero, whereas for $\epsilon_\nu < \mu$ the occupation numbers n_ν are 1. Note that this ρ is

not the physical DM unless $\mu = \epsilon_F$, the Fermi energy. Nevertheless, this argument establishes the existence of a local minimum in a ball around ρ suited to a specific μ. In practice, ϵ_F must be found by varying μ until the correct N is obtained.

It is possible to express our orbital-formulated functional in the language of the DM. Using the lowest-order approximation for the inverse of the overlap matrix, $S^{-1} \approx (2 - S)$, our energy functional takes the form given by Eq. (21). Substituting the expressions for H and S, we obtain

$$\tilde{E} = 2\left[\sum_{\mu\nu} 2R_{\mu\nu}h_{\mu\nu} - (RsR)_{\mu\nu}h_{\mu\nu}\right], \tag{B6}$$

or

$$\tilde{E} = 2\mathrm{Tr}[(2R - RsR)h], \tag{B7}$$

where we made the identification that

$$R_{\mu\nu} = \sum_{i,occ} C^*_{\mu i}C_{\nu i}. \tag{B8}$$

Here, R is a trial DM which coincides with the exact DM ρ only when $\tilde{E}$ is a minimum, or equivalently when the localized orbitals χ_i can be expressed as a linear combination of only the occupied states of the Hamiltonian.

The stability of our orbital procedure is proven in Appendix A. We note that *direct use* of the DM form of the functional [Eq. (B7)] would *not* yield a convergent procedure to estimate the DM, although when implemented in orbital language the method is very robust. The reason for this apparent discrepancy is that *by construction*, the orbital method involves N electrons (the rank of H is N). No such constraint is explicitly built into the functional of Eq. (B7), in which the overlap, Hamiltonian, and R are of dimension M. To see this more clearly, note that the functional derivative of Eq. (B7) gives

$$\frac{\delta\tilde{E}}{\delta R} = [2 - (Rs + sR)]h\ , \tag{B9}$$

which for negative definite h and an orthonormal basis vanishes only for $R = 1$, i.e., a solution producing uniform weighting of *all* the states, which is of course unsuitable for projecting out a part of the spectrum of H or R. In other words, $\tilde{E}$ is stationary for a nonidempotent R. The LNVD functional overcomes this difficulty at the expense of introducing a more complex functional and the introduction of a chemical potential determined *ex post facto*.

The LNVD functional can be derived by using combinations of the possible functionals described in Sec. IV, where we showed that different Taylor polynomial approximations to S^{-1} yield a hierarchy of functionals with minima coinciding with the Kohn-Sham minima. Provided that the Hamiltonian is negative definite and the truncation is made at an odd order, the global minimum of the functional is identical to the Kohn-Sham minimum. It is easy to see that any linear combination of two such polynomial approximations to S^{-1} also leads to a solution if constructed so that the approximation for S^{-1} reduces to the unit matrix when $S = \mathbb{1}$. The first two Taylor approximants to S^{-1} are $\pi_1(S) = \mathbb{1} + (\mathbb{1} - S) = 2\mathbb{1} - S$ and $\pi_2(S) = \mathbb{1} + (\mathbb{1} - S) + (\mathbb{1} - S)^2 = 3\mathbb{1} - 3S + S^2$. The LNVD functional is just $3\pi_1 - 2\pi_2 = 3S - 2S^2$. Of course there are many other such combinations which have the same property, so that the LNVD functional is not unique. It is, however, the polynomial DM functional involving the lowest possible powers of ρ while meeting the necessary stability criteria.

We may view the LNVD method as stemming from an alternate approximation for $S^{-1} \approx 3S - 2S^2$. This is worthwhile, since we have made no assumptions that the basis is orthogonal, and can simply rewrite the functional implied by this form for S^{-1} to obtain a DM functional appropriate for a nonorthogonal basis set. In particular, we have

$$\mathrm{Tr}(S^{-1}H) \approx \mathrm{Tr}[(3S - 2S^2)H]. \tag{B10}$$

Using the definition of R [Eq. (B8)], the definitions of H and S in terms of h and s, and trace invariance, it is easy to see that Eq. (B10) may be rewritten as

$$\mathrm{Tr}[(3RsR - 2RsRsR)h]. \tag{B11}$$

While we have not studied the stability of this functional in detail, we have found it to be stable in simple test cases. The same generalization of the LNVD functional for nonorthonormal basis sets has recently been derived by Nunes and Vanderbilt[63] following a more fundamental approach.

In terms of computational efficiency, there are some advantages to the orbital formulation. (i) The number of electrons is fixed by the construction of our functional. The localized orbitals χ_i may always be expressed as a linear combination of the eigenvectors ψ_i. Our conjugate-gradient search iteratively removes those parts of the χ_i which are associated with eigenvectors conjugate to unoccupied eigenvalues. Thus, so long as the χ_i are linearly independent [so that $\mathrm{rank}(h) = N$], we are bound to compute the correct energy of an N-electron system when we are at the minimum of the functional, so no chemical potential constraint is needed. (ii) An additional advantage of our scheme is the possibility of making a physical guess for the χ_i's. A natural guess in a covalent system is to construct bonds. In c-Si, for example, this yields much better convergence than simply starting with a random guess for the orbitals, which is effectively equivalent to the simple starting guess of $\rho_{\mu\nu} = \delta_{\mu\nu}/2$ used by LNVD. (iii) Finally, this method may be viewed as an effective way to find Wannier functions, which are useful in many ways for constructing physical quantities.

[1] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[2] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992), and references therein.
[3] See the collection of papers in *Solid State Physics: Advances in Research and Applications*, edited by F. Seitz, C. Turnbull, and H. Ehrenreich (Academic Press, New York, 1980), Vol. 35.

[4] R. Haydock, in *Solid State Physics: Advances in Research and Applications* (Ref. 3).
[5] A. R. Williams, P. J. Feibelman, and N. D. Lang, Phys. Rev. B **26**, 5433 (1982).
[6] G. A. Baraff, E. O. Kane, and M. Schluter, Phys. Rev. B **21**, 5662 (1980).
[7] J. Bernholc, N. O. Lipari, and S. T. Pantelides, Phys. Rev. B **21**, 3545 (1980).
[8] O. Gunnarsson, O. Jepsen, and O. K. Andersen, Phys. Rev. B **27**, 7144 (1983).
[9] O. K. Andersen and O. Jepsen, Phys. Rev. Lett. **53**, 2571 (1984).
[10] F. Yndurain, J. D. Joannopoulos, M. L. Cohen, and L. M. Falicov, Solid State Commun. **15**, 617 (1974); J. D. Joannopoulos and F. Yndurain, Phys. Rev. B **10**, 5164 (1974).
[11] S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1992).
[12] W. Zhong, D. Tomanek, and G. F. Bertsch, Solid State Commun. **86**, 607 (1993).
[13] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 1077 (1992).
[14] W. Kohn, Phys. Rev. **115**, 809 (1959).
[15] P. Ordejón, D. A. Drabold, M. P. Grumbach, and R. M. Martin, Phys. Rev. B **48**, 14 646 (1993).
[16] W.- L. Wang and M. Teter, Phys. Rev. B **46**, 12 798 (1992).
[17] F. Mauri, G. Galli, and R. Car, Phys. Rev. B **47**, 9973 (1993).
[18] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).
[19] W. Kohn, Chem. Phys. Lett. **208**, 167 (1993).
[20] E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10 088 (1994).
[21] W. Hierse and E. B. Stechel (unpublished).
[22] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).
[23] D. A. Drabold and O. F. Sankey, Phys. Rev. Lett. **70**, 3631 (1993).
[24] O. F. Sankey, D. A. Drabold, and A. Gibson, Phys. Rev. B **50**, 1376 (1994).
[25] G. H. Golub and C. F. Van Loan, *Matrix Computations* (John Hopkins University Press, Baltimore, 1983).
[26] P. Ordejón, G. Ortiz, and P. Phillips, Phys. Rev. B **50**, 14 682 (1994).
[27] X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).
[28] M. S. Daw, Phys. Rev. B **47**, 10 895 (1993).
[29] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).
[30] R. W. Nunes and D. Vanderbilt, Phys. Rev. Lett. **73**, 712 (1994).
[31] See, for instance, J. L. Mercer and M. Y. Chou, Phys. Rev. B **49**, 8506 (1994); P. Ordejon, D. Lebedenko, and M. Menon, *ibid.* **50**, 5645 (1994).
[32] J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory* (McGraw-Hill, New York, 1970).
[33] P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964).
[34] W. Kohn and L. J. Sham, Phys. Rev. **140**, 1133 (1965).
[35] E. Artacho and L. Milans del Bosch, Phys. Rev. A **43**, 5770 (1991), and references therein.
[36] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, Phys. Rev. Lett. **69**, 1077 (1992).
[37] I. Štich, R. Car, M. Parrinello, and S. Baroni, Phys. Rev. B **39**, 4997 (1989).
[38] N. Troullier, J. R. Chelikowsky, and Y. Saad, Bull. Am. Phys. Soc. **39**, 155 (1994).
[39] S. Satpathy and Z. Pawlowska, Phys. Status Solidi B **145**, 555 (1988), and references therein.
[40] P. Löwdin, J. Chem. Phys. **18**, 365 (1950).
[41] O. F. Sankey and D. J. Niklewski, Phys. Rev. B **40**, 3979 (1989).
[42] J. D. Chadi, Phys. Rev. B **29**, 785 (1984).
[43] This means that the number of electrons of the system is not exactly conserved. The modified charge density, given by Eq. (28), is only equal to the exact value when the states are orthonormal. Otherwise, the expansion of the S^{-1} matrix is not exact, and the charge density at each point $\mathbf{r}$ is smaller than the exact charge density. The error on the total number of electrons of each spin is given by (Ref. 18) $\Delta N = \mathrm{Tr}[(\mathbb{1} - S)^2]$ which is positive for any nonorthonormal set of states ($S \neq \mathbb{1}$). This could be a serious problem for the application of the method to self-consistent Hamiltonians such as the LDA, since small variations in the number of electrons could yield very inaccurate results for the energy. However, the problem can be easily solved by computing the charge density using a higher order in the expansion of the S^{-1} matrix to construct the self-consistent Hamiltonian. The error in the total charge is then reduced to $\Delta N^{(\mathcal{N})} = \mathrm{Tr}[(\mathbb{1} - S)^{\mathcal{N}+1}]$ where $\mathcal{N}$ is the order of the approximation of S^{-1} [Eq. (27)] used to construct the density. For larger $\mathcal{N}$, the calculated charge density approaches the exact value, and the number of electrons can be computed to the necessary degree of accuracy.
[44] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[45] G. B. Bachelet, D. R. Hamann, and M. Schluter, Phys. Rev. B **26**, 4199 (1982).
[46] J. Harris, Phys. Rev. B **31**, 1770 (1985).
[47] D. A. Drabold, P. A. Fedders, and P. Stumm, Phys. Rev. B **49**, 16 415 (1994).
[48] O. F. Sankey and G. B. Adams (unpublished).
[49] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Oxford Science Publications, Clarendon, Oxford, 1987).
[50] J. Broughton and F. Khan, Phys. Rev. B **40**, 12 098 (1989).
[51] G. Pastore, E. Smargiassi, and F. Buda, Phys. Rev. B **44**, 6334 (1991); M. P. Grumbach, D. Hohl, R. M. Martin, and R. Car, J. Phys. Condens. Matter **6**, 1999 (1994).
[52] C. Xu, C. Wang, C. Chan, and K. Ho, J. Phys. Condens. Matter **4**, 6047 (1992).
[53] W. C. Swope, H. C. Andersen, P. H. Berens, and R. K. Wilson, J. Chem. Phys. **76**, 637 (1982).
[54] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, Phys. Rev. B **45**, 1538 (1992).
[55] Note that this is not the case when the wave functions are allowed to delocalize (Ref. 18), in which case all the modes can be made larger than the energy gap (see Appendix A).
[56] P. E. Blöchl and M. Parrinello, Phys. Rev. B **45**, 9413 (1992).
[57] S. Nosé, Mol. Phys. **52**, 255 (1984).
[58] S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[59] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[60] S. Nosé, Mol. Phys. **57**, 187 (1986).
[61] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[62] J. Smith and J. G. Gay, Phys. Rev. B **12**, 4238 (1975).
[63] R. W. Nunes and D. Vanderbilt, Phys. Rev. B **50**, 17 623 (1994).