RAPID COMMUNICATIONS

PHYSICAL REVIEW B VOLUME 47, NUMBER 15 15 APRIL 1993-I

Rapid Communications

Rapid Communications are intended for the accelerated publication of important new results and are therefore given priority treatment both in the editorial office and in production. A Rapid Communication in **Physical Review B** *should be no longer than four printed pages and must be accompanied by an abstract. Page proofs are sent to authors.*

Orbital formulation for electronic-structure calculations with linear system-size scaling

Francesco Mauri, Giulia Galli, and Roberto Car

Institut Romand de Recherche Numérique en Physique des Matériaux (IRRMA), PHB-Ecublens, 1015 Lausanne, Switzerland

(Received 19 January 1993)

A novel energy functional for total-energy and molecular-dynamics calculations is introduced, and proven to have the Kohn-Sham ground-state energy as its absolute minimum. The use of this functional within a localized orbital formulation leads to an algorithm for electronic structure calculations whose computational work load grows linearly with the system size. The foundations and accuracy of the approach and the performances of the algorithm are first discussed analytically and then illustrated with several numerical examples.

I. INTRODUCTION

Total-energy calculations and molecular-dynamics (MD) simulations with forces derived either from first-principles or tight-binding (TB) Hamiltonians form the backbone of many studies of materials performed in condensed-matter physics.[1] However, such investigations have been limited in scale and scope by the computer time required by standard algorithms, which grows as the cube of the system size. This unfavorable scaling has so far precluded the use of first-principles and of TB Hamiltonians for systems involving more than a few hundred and a few thousand electrons, respectively. Furthermore, *ab initio* MD simulations[2] implying times longer than a few picoseconds are still out of reach.

In this paper we present a method for total-energy and MD calculations within density-functional theory. Our scheme is based on an orbital picture and on an energy functional which is proven to have the Kohn-Sham ground-state energy as its absolute minimum. When used within a localized orbital formulation,[3] the method implies an overall computational cost which scales linearly with the system size and can thus be used to tackle a variety of problems so far inaccessible.

Most methods for total-energy calculations rely on an orbital picture, e.g., the Kohn-Sham (KS) formulation of density-functional theory (DFT). Minimization procedures which consider only the density matrix as a variable have been proposed in the literature[4,5] and successfully applied to TB Hamiltonians. However, their generalization to first-principles local-density functional (LDA) calculations[6] appears to be very costly since they imply the knowledge of the full spectrum (occupied and empty states) of the Hamiltonian matrix (H).

The solution of the single-particle eigenvalues problem (KS equations) is usually obtained by diagonalizing H, which is set up according to a chosen basis set for the electronic orbitals $\{\phi\}$. When the number M of basis functions is much larger than the number N of electrons, iterative diagonalization techniques are drastically more efficient than direct diagonalization procedures. However, the overall scaling as a function of the system size of both direct and iterative schemes is usually of $O(N^3)$.

Iterative approaches can be divided into two classes: constrained minimization methods[1] in which the single-particle wave functions are required to be orthonormal and unconstrained (UM) methods,[3,7] in which the orbitals are allowed to overlap. In computations with plane-wave (PW) basis sets and pseudopotentials—which are the ones most widely used in, e.g., first-principles MD simulations—the evaluation of $\{H\phi\}$ costs $O(NM)$ operations (where M is proportional to N), if advantage is taken of fast Fourier transform techniques and of the localized nature of nonlocal pseudopotentials. The application of orthogonality constraints implies instead $O(N^2M)$ operations. When UM are used, the calculation of the overlap matrix ($\mathbf{S}$) and of its inverse are of $O(N^2M)$ and $O(N^3)$, respectively. It has been shown in Ref. 3 that the electronic orbitals can be required to be localized in given regions of space, without any significant loss of accuracy in the calculation. Appropriate localization of orbitals reduces the number of operations needed for $\{H\phi\}$ to $O(N)$. In order to reduce to $O(N)$ also iterative orthogonalization procedures or $\mathbf{S}$ inversion, further assumptions on the form of the overlap matrix are necessary.[8]

A NEW FUNCTIONAL FOR ELECTRONIC STRUCTURE CALCULATIONS

We introduce an unconstrained minimization method in which the inverse of the overlap matrix is replaced by its series expansion in $(\mathbf{I}-\mathbf{S})$ up to an odd order $\mathcal{N}$, where $\mathbf{I}$ is the identity matrix. We first prove that the

total-energy functional defined with $\mathbf{S}^{-1}$ approximated in this way has the Kohn-Sham ground-state energy (E_0) as its absolute minimum.

We consider an energy functional of $N/2$ orbitals $\{\phi\}$ expanded in a finite basis set, and of the $(N/2 \times N/2)$ matrix $\mathbf{A}$:

$$E[\mathbf{A},\{\phi\}] = 2\left(\sum_{ij}^{N/2} A_{ij}\langle\phi_i|-\tfrac{1}{2}\nabla^2|\phi_j\rangle + F[\tilde{\rho}]\right) + \eta\left(N - \int d\mathbf{r}\,\tilde{\rho}(\mathbf{r})\right), \quad (1)$$

where $\tilde{\rho}(\mathbf{r}) = \tilde{\rho}[\mathbf{A},\{\phi\}](\mathbf{r}) = 2\sum_{ij}^{N/2} A_{ij}\phi_j(\mathbf{r})\phi_i(\mathbf{r})$, $F[\tilde{\rho}]$ is the sum of the Hartree, exchange-correlation, and external potential energy functionals, and η a constant to be specified. The factor 2 accounts for the electronic occupation numbers, which are assumed to be all equal. For simplicity we consider real orbitals. If $A_{ij} = S_{ij}^{-1}$, where $S_{ij} = \langle\phi_i|\phi_j\rangle$, then $\tilde{\rho}[\mathbf{S}^{-1}]$ is the single-particle charge density $\rho(\mathbf{r})$ and the term multiplying η is zero; in this case the functional of Eq. (1) is the total energy of interacting electrons in an external field according to DFT, written for overlapping orbitals.[3,7] In particular, if the wave functions are orthonormal (we indicate with $\{\psi\}$ sets of orthonormal orbitals) then $A_{ij} = \delta_{ij}$, and Eq. (1) gives the total-energy functional of DFT ($E^{\perp}[\{\psi\}]$) used in constrained total-energy minimizations and *ab initio* MD simulations.[1] The sets $\{\psi\}$ and $\{\phi\}$ are related by the transformation $\psi_i = \sum_j S_{ij}^{-1/2}\phi_j$ and then $E^{\perp}[\mathbf{S}^{-1/2}\phi] = E[\mathbf{S}^{-1},\{\phi\}]$. Therefore

$$\min_{\{\psi\}} E^{\perp}[\{\psi\}] = \min_{\{\phi\}} E[\mathbf{S}^{-1},\{\phi\}] = E_0. \quad (2)$$

Here we define an energy functional of $\{\phi\}$, $E[\mathbf{Q}[\{\phi\}],\{\phi\}]$, by taking $\mathbf{A} = \mathbf{Q}$ where

$$\mathbf{Q} = \sum_{n=0}^{\mathcal{N}} (\mathbf{I}-\mathbf{S})^n \quad (3)$$

and $\mathcal{N}$ is odd. $\mathbf{Q}$ is the the truncated series expansion of $\mathbf{S}^{-1}$. We note that if the orbitals are orthonormal ($S_{ij} = \delta_{ij}$), $Q_{ij} = \delta_{ij}$ and $E[\mathbf{Q},\{\psi\}]$ coincides with $E^{\perp}[\{\psi\}]$. As a consequence

$$\min_{\{\psi\}} E^{\perp}[\{\psi\}] = \min_{\{\psi\}} E[\mathbf{Q},\{\psi\}] \geq \min_{\{\phi\}} E[\mathbf{Q},\{\phi\}], \quad (4)$$

since $\{\psi\}$ is a subset of $\{\phi\}$.

We now consider the difference between the functionals $E[\mathbf{Q},\{\phi\}]$ and $E[\mathbf{S}^{-1},\{\phi\}]$, i.e.,

$$\Delta E = E[\mathbf{Q},\{\phi\}] - E[\mathbf{S}^{-1},\{\phi\}] = \int_0^1 (\partial E[\mathbf{A}(\lambda),\{\phi\}]/\partial\lambda)\,d\lambda, \quad (5)$$

where $\mathbf{A}(\lambda) = \lambda(\mathbf{Q}-\mathbf{S}^{-1}) + \mathbf{S}^{-1}$. Using Eq. (1), Eq. (5) becomes

$$\Delta E = 2\sum_{ij}^{N/2}\langle\phi_j|-\tfrac{1}{2}\nabla^2 + \overline{V}_{\mathrm{KS}} - \eta|\phi_i\rangle(Q_{ij} - S_{ij}^{-1}), \quad (6)$$

where $\overline{V}_{\mathrm{KS}} = \int_0^1 d\lambda\, V_{\mathrm{KS}}[\tilde{\rho}(\lambda)]$ and $V_{\mathrm{KS}}[\tilde{\rho}] = \frac{\delta F}{\delta\tilde{\rho}}$. The matrix $(\mathbf{Q}-\mathbf{S}^{-1}) = -\mathbf{S}^{-1}(\mathbf{I}-\mathbf{S})^{\mathcal{N}+1} = -(\mathbf{I}-\mathbf{S})^{\mathcal{N}+1}\mathbf{S}^{-1}$ is negative definite, for odd $\mathcal{N}$. Given a finite basis set, one can choose η large enough so that the operator $\overline{H}_{\mathrm{KS}} - \eta = -\frac{1}{2}\nabla^2 + \overline{V}_{\mathrm{KS}} - \eta$ is negative definite; then also the $(N/2 \times N/2)$ matrix $\langle\phi_j|\overline{H}_{\mathrm{KS}} - \eta|\phi_i\rangle$ is negative definite and ΔE is positive since it is equal to the trace of the product of two negative definite matrices. This proves that if η satisfies the above requirement, then for each set of $\{\phi\}$

$$E[\mathbf{Q},\{\phi\}] \geq E[\mathbf{S}^{-1},\{\phi\}], \quad (7)$$

where the equality holds only for $S_{ij} = \delta_{ij}$. From Eqs. (2), (4), and (7) it follows that

$$\min_{\{\psi\}} E^{\perp}[\{\psi\}] = \min_{\{\phi\}} E[\mathbf{Q},\{\phi\}] = \min_{\{\phi\}} E[\mathbf{S}^{-1},\{\phi\}] = E_0 \quad (8)$$

and that the minimization of $E[\mathbf{Q},\{\phi\}]$ yields orthonormal orbitals.

If the Hamiltonian does not depend on ρ, a η larger than the Hamiltonian maximum eigenvalue ensures that $\Delta E \geq 0$. Within LDA, we have $H_{\mathrm{KS}}[\tilde{\rho}] \leq H_H[\rho]$, where $H_H[\rho] = -\frac{1}{2}\nabla^2 + V_H[\rho] + V_{\mathrm{ext}}$, and V_H and V_{ext} are the Hartree and external potential, respectively. This follows from the property $\tilde{\rho}[\mathbf{Q}](\mathbf{r}) \leq \tilde{\rho}[\mathbf{S}^{-1}](\mathbf{r}) = \rho(\mathbf{r})$, valid for each point $\mathbf{r}$, and from the explicit LDA expression of the exchange and correlation energy as a function of $\rho(\mathbf{r})$. Within a PW implementation with a finite cutoff, H_H has an upper bound. This ensures the existence of a η such that $\Delta E \geq 0$.

In practice, for η larger than the highest occupied eigenvalue ϵ_N of $H_{\mathrm{KS}}[\rho_0]$, where ρ_0 is the ground-state charge density, the set of orbitals which minimizes $E^{\perp}[\{\psi\}]$ is a local minimum of $E[\mathbf{Q},\{\phi\}]$. Thus first-principles MD simulations can be performed with $\eta \simeq \epsilon_{N+1}$; this choice of η allows the use of the same time

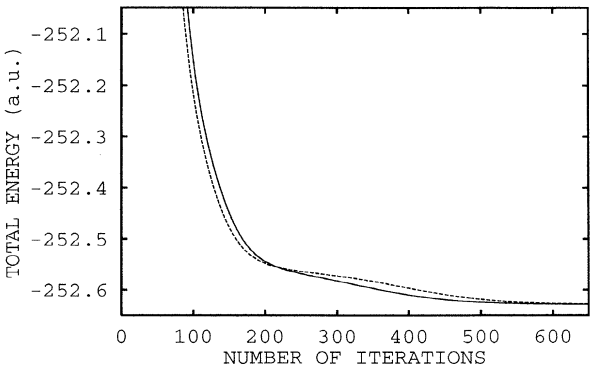

FIG. 1. Total energy as a function of the number of iterations for a steepest descent minimization of 64 Si atoms, described within LDA with PW basis sets. The solid and dotted lines correpond to the minimization of $E[\mathbf{Q}]$ and $E^{\perp}$ (see text), respectively. $\mathbf{Q}$ was defined with $\mathcal{N} = 1$. Kinetic energy cutoffs of 12 and 36 Ry for the wave functions and charge density, respectively, were used in the calculations; each run was started from the same set of random numbers.

step and fictitious mass adopted in standard constrained dynamics for $\mathcal{N}=1$. This can be proved analytically by expanding $E[\mathbf{Q}]$ and $E^{\perp}$ around their minima and calculating the frequencies associated with the electronic degrees of freedom.[8] The optimal value of η for total-energy minimizations depends upon the initial guess used for the $\{\phi\}$ and is usually larger than ϵ_{N+1}.

The functional presented here has clear advantages over standard energy functionals when conjugate and preconditioned conjugate gradient minimization procedures are used: the complication of imposing orthonormality constraints is avoided, and contrary to ordinary unconstrained methods an automatic control of the $\mathbf{S}$ matrix is provided, since at the minimum $S_{ij}=\delta_{ij}$.

Finally we note that our formulation can be related to the density matrix approach adopted in Ref. 5 if the density matrix is constructed similarly to our density $\tilde{\rho}[\mathbf{Q}]$.[8]

We have tested numerically the validity of our formulation for TB (Ref. 9) and KS Hamiltonians. As an example, in Fig. 1 we show the total energy as a function of the number of iterations for a steepest descent minimization of 64 Si atoms, described within LDA with PW basis sets and ordinary pseudopotentials.[10] The calculation was started from orbitals set up from random numbers, with $\eta=3.0$ Ry. The minimizations of $E[\mathbf{Q},\{\phi\}]$ and $E^{\perp}[\{\psi\}]$ required the same number of iterations and lead to the same energy.

LOCALIZED ORBITALS AND AN ALGORITHM WITH LINEAR SYSTEM-SIZE SCALING

We now turn to the discussion of the present method, when used with localized orbitals (LO).[3] Within such a formulation, each single-particle wave function is constrained to be localized in an appropriate region of space [localization region (LR)], i.e., free to vary inside and zero outside the LR. These LR's are centered around different points, for instance, the atomic positions, and their extension does not vary with system size. Different single-particle orbitals can be associated with the same LR (e.g., two orbitals per LR for Si, which has four valence electrons). When LO are used all sums entering the expression of $E[\mathbf{Q}]$ and its derivatives extend only to orbitals belonging to overlapping LR's. It then follows that our method, which does not imply any orthogonalization or $\mathbf{S}$ inversion, leads to an algorithm which scales linearly with the system size.

For a given size of the LR, the minimum of $E[\mathbf{Q}]$ with respect to LO $\{\phi^L\}$ does not coincide with that of $E[\mathbf{S}^{-1}]$, and the LO which minimize $E[\mathbf{Q}]$ in general are not orthonormal. This is easily seen as follows. Whereas Eqs. (4) and (7) still hold, Eq. (2) is no longer valid. Indeed the transformation from $\{\psi\}$ to $\{\phi\}$ with $\mathbf{S}^{-1/2}$ does not preserve the size of the LR (it does not map functions localized in a given region onto functions localized in the *same* region). Therefore Eq. (8) does not hold but is replaced by

$$\min_{\{\psi^L\}} E^{\perp} \geq \min_{\{\phi^L\}} E[\mathbf{Q}] \geq \min_{\{\phi^L\}} E[\mathbf{S}^{-1}] \geq E_0, \qquad (9)$$

where the LR for the $\{\psi^L\}$ and $\{\phi^L\}$ are the same. At the minimum $\mathbf{S}$ is different from $\mathbf{I}$. However, its deviation from $\mathbf{I}$ is limited since the difference $E[\mathbf{Q}]-E[\mathbf{S}^{-1}]$ increases as the $\mathbf{S}$ eigenvalues spread out around 1. The larger $\mathcal{N}$, the wider the spread of $\mathbf{S}$ eigenvalues. The variational quality of the results obtained by minimizing $E[\mathbf{Q}]$, i.e., the difference $[\min_{\{\phi^L\}} E[\mathbf{Q}]-E_0]$, depends upon (i) the order $\mathcal{N}$ chosen for the definition of the $\mathbf{Q}$ matrix and (ii) the size of the LR. For $\mathbf{S}\leq 2\mathbf{I}$ we have $E[\mathbf{Q}(\mathcal{N}-2)]\geq E[\mathbf{Q}(\mathcal{N})]$. Therefore, by increasing $\mathcal{N}$ in the definition of $\mathbf{Q}$, one obtains an improvement of the total energy. This leads as well to an increase of the number of operations needed in the computation of $\mathbf{Q}$ [see Eq. (3)]. Alternatively one may choose to increase the radius of the localization region ($r_c^{\rm loc}$) to improve the quality of the results. We note that the number of nonzero elements of $\mathbf{S}$ is proportional to $n_{\rm LR}N$, where $n_{\rm LR}$ is the average number of regions overlapping with a given one. Instead, the number of degrees of freedom needed to define the $N/2$ single particle orbitals is proportional to mN, where m is the number of points belonging to a LR, e.g., the number of points where the wave function is nonzero. The ratio $n_{\rm LR}/m$ strongly depends on the basis set chosen to set up the Hamiltonian. Therefore, the optimal choice of $\mathcal{N}$ and of $r_c^{\rm loc}$, e.g., of the parameters determining the efficiency and accuracy of the computation, crucially depends upon the chosen basis set.[8]

We note that in calculations where $m\gg n_{\rm LR}$, the computer time for the $\mathbf{S}$ inversion amounts to a small fraction of the total time also for relatively large systems (e.g., systems with up to a few thousand electrons in LDA calculations with PW basis). On the other hand, for computations with small basis sets, such as those with TB Hamiltonians, the computer time for the $\mathbf{S}$ inversion constitutes a considerable part of the total time also for small systems.

NUMERICAL RESULTS

We have tested the formulation with LO for TB Hamiltonians[9] with $\varepsilon_s+\varepsilon_p=0$. Table I shows the cohesive energy E_c of Si in the diamond structure, computed with a 216-atom supercell and simple-cubic periodic boundary conditions; E_c has been evaluated with

TABLE I. Cohesive energy of Si computed with a 216-atom supercell, with $\mathbf{Q}[\mathcal{N}=1]$ and $\mathbf{Q}[\mathcal{N}=3]$ [see Eq. (3)] and for different choices of η (eV) [see Eq. (1)], as a function of the number of neighbor shells (N_s) included in the definition of the localization region.

N_s	$\eta=8$ $\mathcal{N}=1$	$\eta=3$ $\mathcal{N}=1$	$\eta=3$ $\mathcal{N}=3$
1	4.4676	4.6066	5.0723
2	5.3063	5.3289	5.3985
3	5.3402	5.3644	5.4179
4	5.3449	5.3683	5.4187
5	5.4006	5.4102	5.4352
∞	5.4440	5.4440	5.4440

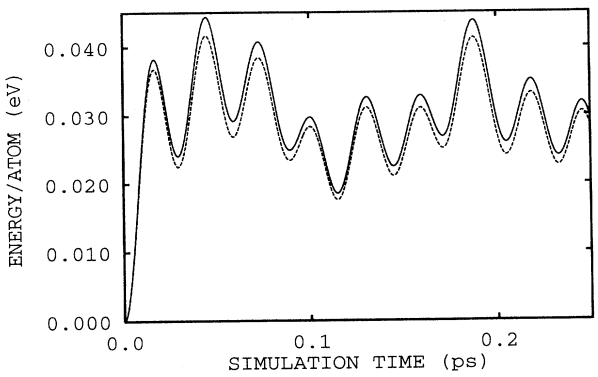

FIG. 2. Oscillations of the total energy of 64 Si atoms measured over a MD run of 0.25 ps (see text). The system has been described with a TB Hamiltonian (Ref. 9). The results of a Car-Parrinello (CP) MD dynamics using the functional $E[\mathbf{Q}]$ with $\mathcal{N} = 1$, $\eta = 3$ eV and LR's defined by $N_s = 2$ (solid line) are compared to those of exact diagonalization of the Hamiltonian at each atomic position (dotted line), where no localization of orbitals was imposed. In the CP dynamics we used a time step equal to 10 a.u. and $\mu = 300$ a.u.

$\mathbf{Q}[\mathcal{N} = 1]$ and $\mathbf{Q}[\mathcal{N} = 3]$ and for different choices of η, as a function of the size of the LR. In all cases $E[\mathbf{Q}]$ has been minimized with a conjugate gradient procedure, starting from random numbers, and only the Γ point has been included in the supercell Brillouin-zone sampling. The LR's have been centered around atomic sites and two orbitals were assigned to each LR. It is seen that E_c converges rapidly as a function of the number of neighbor shells (N_s) included in the definition of the LR, with both $\mathcal{N} = 1$ and 3. Already with $N_s = 2$ we obtain very good results, i.e., E_c higher than the exact result by only 2.1% and 0.8% for $\mathcal{N} = 1$ and 3, respectively.

We have explicitly verified the linear system size scaling of the algorithm by computing E_c with 64-, 216-, and 1000-atom supercells, starting from random atomic orbitals, with $\eta = 8$ eV, $\mathcal{N} = 1$ and $N_s = 2$. The CPU time per step scales linearly with N, the number of iterations needed to converge E_c up to the fifth significant digit does not vary with N and it is equal to 80.

In order to test the method for MD simulations, we have performed a MD run for 64 Si atoms, starting from the diamond lattice equilibrium positions, with random velocities corresponding to a temperature of about 400 K. Figure 2 shows the oscillations of the total potential energy measured over 0.25 ps, computed by solving the coupled equations of motion[2] $\mu\ddot{\phi}_i = -\frac{\delta E[\mathbf{Q}]}{\delta \phi_i}$ and $M_I\ddot{\mathbf{R}}_I = -\nabla_I E[\mathbf{Q}]$ for the electronic and ionic ($\mathbf{R}_I$) degrees of freedom, respectively. M_I are the ionic masses and μ is the fictitious electronic mass.[2] $\mathbf{Q}$ has been defined with $\mathcal{N} = 1$ and the LR with N_s=2. The results obtained in such a way (solid line) are compared to those of exact diagonalization of the Hamiltonian at each atomic position, with no localization of orbitals (dotted line). The agreement between the two calculations is excellent, the difference between the two curves of Fig. 2 being of the order of meV/atom. The same numerical tests presented for Si were carried out also for C, yielding very similar results.

CONCLUSIONS

We have presented an energy functional for total-energy and molecular-dynamics calculations which has the Kohn-Sham ground-state energy as its absolute minimum. A crucial feature of this functional is that its minimization does not imply either explicit orthogonalization of the orbitals or inversion of an overlap matrix. The calculation is highly stable from a numerical point of view with respect to UM procedures, since the overlap matrix is kept close to unity and then automatically controlled. The use of this approach within a localized orbital formulation[3] leads straightforwardly to a method whose computational work load grows linearly with the system size. The performances and efficiency of the method have been illustrated with several numerical examples for semiconducting systems. In particular we have presented the first MD simulation with localized orbitals which show (i) the feasibility of calculations with LO for nonsymmetric systems and (ii) that such simulations can be performed with the same parameters (time step and fictitious electronic mass) as those used in standard Car-Parrinello-like dynamics. Numerical analysis for metallic systems are underway.

ACKNOWLEDGMENTS

We thank S. de Gironcoli, F. Gygi, and R. M. Martin for useful discussions. We acknowledge support by the Swiss National Science Foundation under Grant No. 21-31144.91.

[1]For a review see, e.g., G. Galli and A. Pasquarello, *New Perspectives on Computer Simulation in Chemical Physics* (Kluwer, Dordrecht, in press).
[2]R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[3]G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[4]R. Haydock, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1980), Vol. 35, p. 215.
[5]X.-P. Li, W. Nunes, and D. Vanderbilt (unpublished).
[6]S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1991).
[7]M. C. Arias, T. A. Payne, and J. D. Joannopoulos, Phys. Rev. Lett. **69**, 1077 (1992).
[8]F. Mauri and G. Galli (unpublished).
[9]L. Goodwin, A. J. Skinner, and D. G. Pettifor, Europhys. Lett. **9**, 701 (1989).
[10]G. B. Bachelet, D. R. Hamann, and M. Schlueter, Phys. Rev. B **26**, 4199 (1982).