Efficient pseudopotentials for plane-wave calculations. II. Operators for fast iterative diagonalization

N. Troullier and José Luís Martins
Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455
(Received 15 October 1990; revised manuscript received 4 January 1991)

We present an investigation of the computational requirements for the pseudopotential plane-wave method as a function of the number of atoms per unit cell. For systems containing a large number of atoms the computational load can be reduced if the pseudopotential operator is of a suitable form, such that it can be efficiently calculated in the position representation. The pseudopotentials examined here include local pseudopotentials, position-dependent electron-mass pseudopotentials, and separable nonlocal pseudopotentials.

I. INTRODUCTION

It is one of the goals of electronic-structure theory to be able to perform calculations for crystals with a large number of atoms per unit cell. For supercell simulations of defects, surfaces, alloys, amorphous solids, or liquids, the number of atoms can be as large as 50 to 100 atoms. The different algorithms used for these types of calculations have varying dependencies for the computational requirements on the number of atoms in the unit cell. Therefore, algorithms that are the most computationally efficient for one to eight atoms per unit cell may not be the most efficient for, say, 50 to 200 atoms.

In modern pseudopotential codes using iterative diagonalization algorithms[1-4] the crucial computational step is the calculation of the product $H\Psi$, where H is the Hamiltonian and Ψ is a trial wave vector. If H is represented in a basis set of N functions, $O(N^2)$ words are required to store H, and $O(N^2)$ operations are required to calculate $H\Psi$ for each trial eigenvector.[5] For a general form of the Hamiltonian it is not possible to avoid this scaling law, but for particular forms of H there are cases where one can do better. If H is a sparse matrix, both storage and operations may be proportional to N. Examples of sparse matrices in electronic-structure theory are the kinetic-energy operator in a plane-wave basis, the Hamiltonian in a finite element basis, or in any other localized basis set with finite range functions. Also, if $H\Psi$ is a convolution (i.e., H is a generalized Toeplitz matrix), then the storage requirements are $O(N)$ and the number of operations are $O(N \ln N)$. An example of a convolution is the product of a local potential by a wave vector in a plane-wave basis set.

In a plane-wave basis set the kinetic-energy operator $\hat{K}$ is diagonal, and $\hat{K}\Psi$ can be calculated with $O(N)$ operations. The local potential operator $\hat{V}_L$ includes both the local part of the pseudopotential and the screening potential. The product $\hat{V}_L\Psi$ in a plane-wave basis-set momentum representation is a convolution, or, in other words, $\hat{V}_L$ is diagonal in the position representation. This convolution can be calculated using fast Fourier transforms[1,2] (FFT) requiring only $O(N \ln N)$ operations to evaluate $\hat{V}_L\Psi$. In the absence of a similar computational trick, the product $\hat{V}_{NL}\Psi$ of the nonlocal part of the pseudopotential by a trial wave vector has to be calculated as a matrix by vector product and therefore requires $O(N^2)$ operations to be performed, and $O(N^2)$ words of memory to store the matrix elements of $\hat{V}_{NL}$. For a constant atomic density and similar accuracy of the basis set, the number of wave functions in the basis set is proportional to $N = n' n_{\text{at.}}$, where $n_{\text{at.}}$ is the number of atoms in the unit cell and the proportionality factor n' is the number of basis wave functions per atom. We see that in the general case for the calculation of $H\Psi = (\hat{K} + \hat{V}_L + \hat{V}_{\text{NL}})\Psi$ in a plane-wave basis for large numbers of atoms, the computing requirements to calculate $\hat{K}\Psi$, $\hat{V}_L\Psi$, and $\hat{V}_{NL}\Psi$ are, respectively, proportional to $n_{\text{at.}}$, $n_{\text{at.}} \ln n_{\text{at.}}$, and $n_{\text{at.}}^2$.

In a previous paper[6] we discussed how one can minimize the size n' of the plane-wave expansion without compromising the accuracy of the calculation. Here we present a discussion of three possible means of modifying the nonlocal pseudopotential operator to improve upon the quadratic dependence on $n_{\text{at.}}$ in the calculation of $\hat{V}_{NL}\Psi$. The first solution is a trivial one; we make the pseudopotential local, i.e., $\hat{V}_{NL} = \hat{0}$. This can only be achieved by making the local pseudopotential harder[7] and therefore is only efficient when $n_{\text{at.}}$ is larger than a threshold value. The second is to use a position-dependent electron-mass pseudopotential introduced by Bachelet *et al.* in the context of quantum Monte Carlo calculations.[8] The local and "differential" character of this operator allows it to be efficiently calculated using FFT's.[8,9] The third approach is to extend the dual-space formalism of Martins and Cohen[1] to nonlocal pseudopotentials.[10,11] We show that this is particularly simple for a separable pseudopotential.[12,13] A general insight into optimizing the operator formalism is given in Sec. II. Section III discusses the use of local pseudopotentials and describes a variational generation method using silicon as an example. Modifying the nonlocal effect into a position-dependent electron-mass operator is the topic of Sec. IV. In Sec. V the application of the dual-space formalism to nonlocal pseudopotentials in real space is dis-

cussed. Atomic units are used throughout this paper unless otherwise indicated. Although this paper is self-contained, the reader may find useful the discussion of pseudopotential generation schemes in Ref. 6, which is a companion to the present paper.

II. PSEUDOPOTENTIAL PLANE-WAVE FORMALISM

In the pseudopotential plane-wave formalism,[14] and using the local-density approximation[15] (LDA), the wave function $\Psi_{\mathbf{k}}(\mathbf{r})$ is expanded into a Fourier series,

$$\Psi_{\mathbf{k}}(\mathbf{r})=\sum_j \Psi_{\mathbf{k}}(\mathbf{G}_j)e^{i(\mathbf{G}_j+\mathbf{k})\cdot\mathbf{r}} , \tag{1}$$

where the $\Psi_{\mathbf{k}}(\mathbf{G}_j)$ are the coefficients of the expansion.[16] In practice, the expansion is truncated and only plane waves with a kinetic energy that is less than a chosen cutoff, $\frac{1}{2}(\mathbf{G}_j+\mathbf{k})^2 \le E_{\text{cut}}$, are included. In this basis set the Schrödinger equation becomes a linear eigenvalue equation,

$$\sum_j H_{ij}(\mathbf{k})\Psi_{\mathbf{k}}(\mathbf{G}_j)=\varepsilon_{\mathbf{k}}\Psi_{\mathbf{k}}(\mathbf{G}_i) , \tag{2}$$

where $H_{ij}(\mathbf{k})$ is the momentum space representation of the Hamiltonian matrix for the point $\mathbf{k}$ in the Brillouin zone. The Hamiltonian matrix is composed of three operators,

$$H_{ij}(\mathbf{k})=\tfrac{1}{2}\delta_{ij}|\mathbf{G}_i+\mathbf{k}|^2+V_L(\mathbf{G}_i-\mathbf{G}_j) + \sum_l V_{NL}^l(\mathbf{G}_i+\mathbf{k},\mathbf{G}_j+\mathbf{k}) , \tag{3}$$

where $\frac{1}{2}\delta_{ij}|\mathbf{q}|^2$ is the kinetic-energy operator $\hat{K}$; $V_L(\mathbf{q})$ is the Fourier-transformed local operator $\hat{V}_L$, which includes the local pseudopotential $V_{\text{local}}(r)$ and the self-consistent electron screening potential; and $\sum_l V_{NL}^l(\mathbf{q},\mathbf{q}')$ is the sum of the momentum space representations of the angular momentum channels of the nonlocal pseudopotential operator $\hat{V}_{NL}$.

Although the methods of solution of this type of eigenvalue problem are well established, general-purpose algorithms may require an enormous amount of computer time if the basis set is large. For a matrix of size N, the number of operations for a Givens-Householder procedure[17] scale as N^3, thereby making the procedure scale as the cube of the number of atoms in the unit cell. Since we are only interested in the M lowest eigensolutions and M is usually much smaller than N, iterative methods may be more efficient because they require only $O(MN^2)$ operations for a general matrix. If the matrix has a special structure such that the product $H\Psi$ can be calculated in less than $O(N^2)$ operations, then the iterative method will be even faster.[18] In this paper we will present three forms of the pseudopotential operator that allows a fast calculation of the product $H\Psi$ in a plane-wave basis.

Let us consider now in more detail the three terms of the crystal Hamiltonian in the pseudopotential plane-wave method [Eq. (3)] and see how efficiently we can calculate their product with a trial wave function. The kinetic-energy operator $\hat{K}=\frac{1}{2}\delta_{ij}|\mathbf{G}_i+\mathbf{k}|^2$ is diagonal in the momentum space representation and the operation of it with the wave vector Ψ is trivial. If the number of plane waves in the basis set is $N=n_{\text{at.}}n'$, then $\hat{K}$ can be stored in $O(n_{\text{at.}}n')$ words and $\hat{K}\Psi$ can be calculated with $O(n_{\text{at.}}n')$ operations. We recall that n', the number of plane waves per atom, depends on the energy cutoff E_{cut} and the volume per atom, but not on the number of atoms per unit cell. If we determine the local potential interaction

$$\hat{V}_L\Psi(\mathbf{G}_j)=\sum_i V_L(\mathbf{G}_i-\mathbf{G}_j)\Psi_{\mathbf{k}}(\mathbf{G}_i)$$

as a matrix vector multiplication, we would need $O(n_{\text{at.}}^2 n'^2)$ words to store the potential and $O(n_{\text{at.}}^2 n'^2)$ operations to calculate the interaction. However, if we consider $\hat{V}_L\Psi$ as a convolution, then we can calculate the convolution in $O(n_{\text{at.}}n'\ln(n_{\text{at.}}n'))$ operations[17] using fast Fourier transforms (FFT) and we need only $O(n_{\text{at.}}n')$ words to store the potential. Physically, calculating $\hat{V}_L\Psi$ as a convolution corresponds to Fourier transforming Ψ to real space where $\hat{V}_L$ is a diagonal operator, multiplying the local potential by the wave function on a real-space grid, and Fourier transforming the product back to the reciprocal space representation.[1,2]

For the semilocal form (nonlocal in angular coordinates but local in the radial coordinate) of the nonlocal component of the pseudopotential [Eq. (3)], the nonlocal potential matrix elements are[14]

$$V_{NL}^l(\mathbf{q},\mathbf{q}')=\frac{1}{\Omega}\int_0^\infty V_{NL,l}(r)j_l(qr)j_l(q'r)r^2dr \times \sum_{m=-l}^{l} Y_{lm}(\hat{\mathbf{q}})Y_{lm}^*(\hat{\mathbf{q}}') , \tag{4}$$

where Ω is the cell volume, $V_{NL,l}(r)$ is the nonlocal part of the pseudopotential, $j_l(qr)$ are the spherical Bessel functions, and $Y_{lm}(\hat{\mathbf{q}})$ are the spherical harmonics which depend only on the direction of the argument. In this matrix form, the nonlocal potential elements require the storage of $O(n_{\text{at.}}^2 n'^2)$ words and $O(n_{\text{at.}}^2 n'^2)$ operations to calculate

$$\hat{V}_{NL}\Psi(\mathbf{G}_j)=\sum_i \left[\sum_l V_{NL}^l(\mathbf{G}_i+\mathbf{k},\mathbf{G}_j+\mathbf{k})\right]\Psi_{\mathbf{k}}(\mathbf{G}_i) .$$

From this we can see that in the calculation of $H\Psi$ for large unit cells, the operation of the nonlocal potential becomes the limiting performance step because it has the strongest (quadratic) dependence on $n_{\text{at.}}$ of the three operators.

III. LOCAL PSEUDOPOTENTIALS

The trivial way to avoid the quadratic scaling with $n_{\text{at.}}$ in the calculation of $H\Psi$ is to use a local pseudopotential. Starkloff and Joannopoulos[7] have shown that a good local pseudopotential could be constructed for Te by making the pseudopotential sufficiently hard. A pseudopotential that is hard requires a larger number n' of plane waves per atom and therefore will only be more computationally efficient than a softer nonlocal pseudopotential when $n_{\text{at.}}$ is sufficiently large. Starkloff and Joannopoulos[7]

start by defining a "bare" potential by subtracting the valence screening from the screened all-electron potential. This bare potential is then multiplied by the steplike function

$$(1+e^{-\lambda r})/(1+e^{-\lambda(r-r_c)})$$

to generate the pseudopotential. The parameters λ and r_c are obtained by fitting the pseudoeigenvalues ε_l^{PP} to the all-electron eigenvalues ε_l^{AE}. Our procedure is similar, but we will have a larger variational freedom, fit both to the eigenvalues and wave functions, and include a bias for smooth pseudopotentials.

In this section and in Sec. IV we will use variational procedures to generate pseudopotentials. In these cases, the norm-conserving conditions[19] are only approximately satisfied, and the use of a parametrized screened pseudopotential as the variational quantity would cause a problem. This arises, because the variational pseudo-wave-functions $R_l^{PP}(r)$ do not match exactly the all-electron wave functions $R_l^{AE}(r)$ outside of the chosen cutoff radius r_{cl}. Since the wave functions do not match, the charge density for the valence all-electron wave functions, $\rho_{\rm val}^{AE}(r)$, will not match the charge density from the pseudo-wave-functions, $\rho^{PP}(r)$ in this region. This causes the valence all-electron Hartree term $V_H[\rho_{\rm val}^{AE};r]$ and the valence all-electron exchange-correlation term $V_{\rm xc}(\rho_{\rm val}^{AE}(r))$ to not match exactly the values $V_H[\rho^{PP};r]$ and $V_{\rm xc}(\rho^{PP}(r))$, obtained from the pseudo-wave-functions in this region. The result of unscreening with these not quite "correct" pseudo-wave-functions will be a non-Coulombic tail in the ionic pseudopotential $V_{\rm ion}(r)$ outside the cutoff radii r_{cl}. This is undesirable and we prefer to use the "unscreened" ionic potential as the function to be varied, and calculate the corresponding pseudo-wave-functions and pseudoeigenvalues self-consistently.

In the pseudopotential variational generation procedure we start by solving the self-consistent all-electron problem to find the eigenvalues ε_l^{AE}, wave functions $R_l^{AE}(r)$, and self-consistent electron screening potential $V_{H+\rm xc}[\rho^{AE};r]$, for a chosen reference atomic configuration. Using only the valence all-electron wave functions, we calculate a valence electron screening potential $V_{H+\rm xc}[\rho_{\rm val}^{AE};r]$. Using the total screening potential $V_{H+\rm xc}[\rho^{AE};r]$, the valence electron screening potential $V_{H+\rm xc}[\rho_{\rm val}^{AE};r]$ and the nuclear charge Z, we calculate a bare ionic potential,

$$V_{\rm bare}(r)=-\frac{Z}{r}+V_{H+\rm xc}[\rho^{AE};r]-V_{H+\rm xc}[\rho_{\rm val}^{AE};r]\ . \tag{5}$$

At this point we defined our ionic local pseudopotential as

$$V_{\rm ion}(r)=\begin{cases} V_{\rm bare}(r) & \text{if } r\geq r_c \\ \sum_{i=0}^{k} v_i r^i & \text{if } r<r_c\ , \end{cases} \tag{6}$$

where r_c is the chosen cutoff radius and the v_i's are the variational parameters. We will require that the ionic pseudopotential be continuous and its first two derivatives be continuous at the cutoff radius r_c and that the ionic pseudopotential does not have a cusp at the origin:

$$\begin{aligned} V_{\rm ion}(r_c)&=V_{\rm bare}(r_c)\ , \\ V'_{\rm ion}(r_c)&=V'_{\rm bare}(r_c)\ , \\ V''_{\rm ion}(r_c)&=V''_{\rm bare}(r_c)\ , \\ V'_{\rm ion}(0)&=0\ . \end{aligned} \tag{7}$$

These conditions were intended to smooth any kink at the cutoff radii and cusp at the origin which would produce a slowly decaying behavior in the Fourier transform of the pseudopotential. These constraints reduce the number of degrees of freedom for the pseudopotential by 4. The other remaining degrees are determined by minimizing a function that evaluates how close a trial pseudopotential satisfies the norm-conserving conditions[19] and the hardness of the pseudopotential. This function is the sum of three types of functions,

$$F(v_0,\ldots,v_k)=\sum_l Q_{{\rm eig},l}+\sum_l Q_{{\rm charge},l}+Q_{FT}\ . \tag{8}$$

The function $Q_{{\rm eig},l}$ measures how accurate the eigenvalues are for each angular momentum l by the function

$$Q_{{\rm eig},l}=\begin{cases} \Delta\varepsilon_l^2 & \text{if } \Delta\varepsilon_l\leq 200 \text{ meV} \\ 100\Delta\varepsilon_l^2-3.96 & \text{if } \Delta\varepsilon_l>200 \text{ meV}\ , \end{cases} \tag{9}$$

where $\Delta\varepsilon_l=\varepsilon_L^{AE}-\varepsilon_l^{PP}$. Norm conservation for each angular momentum is measured by the second function

$$Q_{{\rm charge},l}=\begin{cases} 10I_l^2 & \text{if } |I_l|\leq 0.002 \\ 1000I_l^2-39.6 & \text{if } |I_l|>0.002\ , \end{cases} \tag{10}$$

where I_l is determined by

$$I_l=\frac{\int_0^{r_c}|R_l^{AE}(r)|^2r^2dr-\int_0^{r_c}|R_l^{PP}(r)|^2r^2dr}{\int_0^{r_c}|R_l^{AE}(r)|^2r^2dr}\ . \tag{11}$$

The formula for these errors was designed to be double parabolic in nature. When the errors are within the accepted limits, 200 meV and 0.002, only a small weight is applied. But outside these limits, the weight is increased by two orders of magnitude, indicating that the errors are unacceptable. The last function Q_{FT} controls the hardness of the pseudopotential and has the form

$$Q_{FT}=1.3\int_0^{\infty}|V_{\rm scr}(q)|q^2dq\ , \tag{12}$$

where $V_{\rm scr}(q)$ is the Fourier transform of the screened local potential. The numerical constants used in Eqs. (9), (10), and (12) were determined by trial and error for a silicon pseudopotential with a reasonable total-energy convergence within 24 Ry. We expect that different values would be appropriate for other elements and total-energy convergence requirements.

Direct variation of the polynomial coefficients is highly ill conditioned and not advisable for use in a minimization algorithm. In order to remedy this, the remaining $k-3$ degrees of freedom in Eq. (6) were fixed by directly

varying the ionic potential at the points $r=0$, $r=[1/(k-3)]r_c, \ldots,$ and $r=[(k-4)/(k-3)]r_c$. From these values of the ionic pseudopotential and the boundary conditions [Eq. (7)] the coefficients of the polynomial expansion are found by solving a set of k coupled linear equations. For the order of the polynomial we used $k=7$. Increasing the order up to $k=11$ gave only marginal improvements.

We used silicon to test the ability of the local pseudopotential to adequately reproduce the nonlocal effects. The reference configuration used was $s^2p^{0.5}d^{0.5}$, the cutoff radius was $r_c=2.25a_0$, and the exchange correlation used was Ceperley-Alder,[20] as parametrized by Perdew and Zunger.[21] In Fig. 1 we show the ionic pseudopotential and its Fourier transform, generated using this variational method. We then used this pseudopotential to calculate the structural properties of silicon. The calculations were carried out using an energy cutoff of $E_{\rm cut}=24$ Ry and 10 **k** points in the irreducible wedge of the Brillouin zone. In Table I, column 1, some calculated material properties obtained from a fit of these calculations to the Murnaghan[22] equation of state are listed. Looking at the values in the table, we can see that it is descriptive of silicon, but the band gap and valence-band width as compared to the more accurate nonlocal pseudopotential calculations shows some discrepancy. Upon generating a local pseudopotential for germanium and other materials, we found similar small disagreements.

Trying to improve upon the pseudopotential transferability, we generated a silicon local pseudopotential with

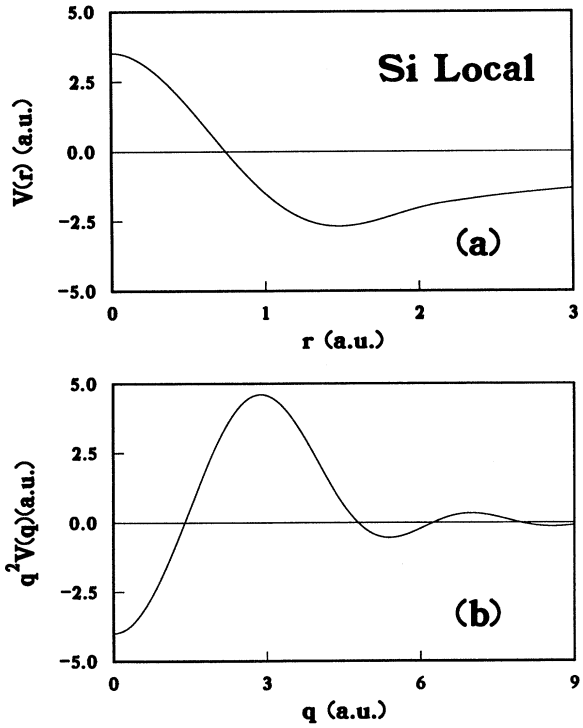

FIG. 1. The local silicon ionic pseudopotential (a) and its Fourier transform (b) obtained from the variational minimization procedure given in Sec. III.

TABLE I. A comparison of some properties of Si obtained from local, position-dependent electron-mass (PDEM) and Kleinman and Bylander (KB) separable nonlocal pseudopotential calculations. The experimental values are also given. The theoretical values of the lattice constant a_0, bulk modulus B_0, pressure derivative of the bulk modulus B_0', and the cohesive energy E_c are obtained from a fit to the Murnaghan (Ref. 22) equation of state. E_g is the band gap and $\Gamma_{25}'-\Gamma_1$ is the total valence bandwidth.

	Local	PDEM	KB	Experiment
a_0 (Å)	5.52	5.48	5.39	5.43[a]
B_0 (GPa)	92.3	91.2	98.4	98.8[a]
B_0'	4.04	4.02	4.10	4.09[b]
E_c (eV)	5.28	5.16	5.28	4.63[a]
E_g (eV)	0.44	0.50	0.51	1.17[a]
$\Gamma_{25}'-\Gamma_1$ (eV)	12.65	12.24	12.04	12.4[c]

[a]Experimental values quoted by C. Kittel, *Introduction to Solid State Physics*, 6th ed. (Wiley, New York, 1986).
[b]A. G. Beattie and J. E. Schirber, Phys. Rev. B **1**, 1548 (1970).
[c]W. D. Grobman and D. E. Eastman, Phys. Rev. Lett. **29**, 1508 (1972).

errors in the eigenvalues and norm conservation smaller than 100 meV and 0.001, respectively. While this pseudopotential had indeed an improved transferability to the solid, it proved to be very hard and required a 36-Ry plane-wave cutoff energy before the total energy was reasonably converged. Therefore, to make an efficient local pseudopotential, a compromise must be made between its hardness and its transferability. As the local silicon pseudopotential is made more accurate, it becomes harder, and the number of atoms at which it becomes more efficient than a nonlocal pseudopotential becomes larger. In the case of the hard ($E_{\rm cut}=36$ Ry) local silicon pseudopotential, this would occur at ~ 50 atoms per unit cell with our current implementation of the plane-wave method. If only general chemical trends are to be determined in a large-scale calculation, then a softer local pseudopotential is appropriate. If, on the other hand, small energy differences must be determined, the local pseudopotential must be made very hard. This means that the threshold on the number of atoms where the local pseudopotential is more efficient than a nonlocal pseudopotential depends on the compromise we are ready to accept in the transferability of the local pseudopotential.

IV. POSITION-DEPENDENT ELECTRON-MASS PSEUDOPOTENTIALS

Reformatting the nonlocal angular-momentum-dependent pseudopotential into a local, yet still angular-dependent, operator for use in a Green's-function Monte Carlo method was the topic of a recent paper by Bachelet, Ceperley, and Chiocchetti.[8] They constructed an atomic pseudo-Hamiltonian of the form

$$H^{PP}(r)=-\tfrac{1}{2}\nabla^2-\tfrac{1}{2}\nabla a(r)\nabla+\frac{b(r)\mathbf{L}^2}{2r^2}+V_L(r)\ , \quad (13)$$

where $a(r)$, $b(r)$, and $V_L(r)$ are radial functions and **L** is

the angular momentum operator, and the functions $a(r)$ and $b(r)$ are constrained by the two conditions

$$1+a(r)\geq 0 \quad \text{and} \quad 1+a(r)+b(r)\geq 0 , \tag{14}$$

in order for the Hamiltonian to remain bounded from below. It has been pointed out that there are some inherent problems in the generation of these types of pseudopotentials for certain elements, among them the transition metals and first row elements.[9]

Both of the two new operators can be efficiently calculated using FFT's because they are local in real space.[9] In Fourier space we have for the second operator in Eq. (13),

$$\nabla\cdot a(r)I\cdot\nabla \Longrightarrow -(\mathbf{G}_j+\mathbf{k})\cdot a(|\mathbf{G}_j-\mathbf{G}_i|)I\cdot(\mathbf{G}_i+\mathbf{k}) , \tag{15}$$

where I is the identity matrix. With this operator, a trial wave function is multiplied by $(\mathbf{G}_i+\mathbf{k})$, giving the three components of its gradient in the momentum space representation. Using FFT's, the three components are obtained on a real-space grid where they are then multiplied by the function $a(r)$. FFT's are then again used to change the product back to momentum space where the dot product of $(\mathbf{G}_j+\mathbf{k})$ with $a(r)\nabla\Psi$ is calculated. The result is then $\nabla\cdot a(r)I\cdot\nabla\Psi$ in the momentum space representation. The computational time required to handle this operator using FFT's is only about three times what is required to handle the local potential.

Using the definition of the angular momentum operator and after some tedious algebra, the third operator of Eq. (13) can be written as

$$\frac{b(r)}{2r^2}\mathbf{L}^2=\nabla\cdot\frac{b(r)}{2r^2}(r^2I-\mathbf{r}\mathbf{r}^T)\cdot\nabla , \tag{16}$$

where we consider $\mathbf{r}$ as a column vector. We have now the additional overhead of finding the 3×3 matrices, $r^2I-\mathbf{r}\mathbf{r}^T$, for each grid point of the FFT sampling in real space, and for each atom in the unit cell. After summing over all atoms in the unit cell, we find out that the contribution of the second and third terms of Eq. (13) can be written as $\nabla\cdot A(\mathbf{r})\cdot\nabla$, where $A(\mathbf{r})$ is a 3×3 matrix. As the bulk of the computation is still the calculation of the FFT's of the three components of the gradient, this entire procedure would still take approximately three times the computer time required to handle the local potential.

In this work we tested only the inclusion of the function $a(r)$ into the formalism as it is trivial to implement, requiring only a few extra lines of computer code. With the additional degrees of freedom provided by $a(r)$, we obtained an improvement in pseudopotential transferability over the local pseudopotential described in Sec. III for the same convergence of the total energy.

To generate these position-dependent electron-mass pseudopotentials (PDEM), we followed the same general procedure described in Sec. III, but with the following additions: The function $a(r)$ was constructed using the constraints

$$\begin{aligned} &a(r_c)=0.0, \quad a'(r_c)=0.0 , \\ &a''(r_c)=0.0, \quad a'(0)=0.0 . \end{aligned} \tag{17}$$

The Fourier transform of the function $a(r)$ was used as a smoothness estimator,

$$Q_{FT:a}=7.5\int_0^\infty |a(q)|q^3dq , \tag{18}$$

which was added to Eq. (8). As an initial starting point, the function $a(r)$ was set to zero. To handle the constraint of Eq. (14), whenever the lower bound of $a(r)$ was smaller than -1.0, an extremely large positive quantity was added to Eq. (8). We also increased the value of the prefactor for the local Fourier transform weight to 47 [Eq. (12)] and decreased the limits on the acceptable error for the eigenvalues and norm conservation to 100 meV and 0.001 [Eqs. (9) and (10)], respectively. Again these coefficients were found by trial and error for silicon, and may be different for other elements.

In our silicon test case, we used the same reference configuration and exchange correlation as we had in Sec. III, but with a smaller cutoff radius $r_c=2.05a_0$. Using this combination, a pseudopotential was generated (see Fig. 2) that had a reasonable converged total energy at a plane-wave cutoff energy of $E_{\text{cut}}=20$ Ry. We then calculated the structural properties of silicon using $E_{\text{cut}}=20$ Ry, and 10 $\mathbf{k}$ points in the irreducible wedge of the Brillouin zone. In Table I, column 2, the material properties obtain from a Murnaghan[22] equation of state fit to the calculated data are listed. From the table we can see that there is a slight improvement over the local pseudopotential values for describing silicon, but there still is a small

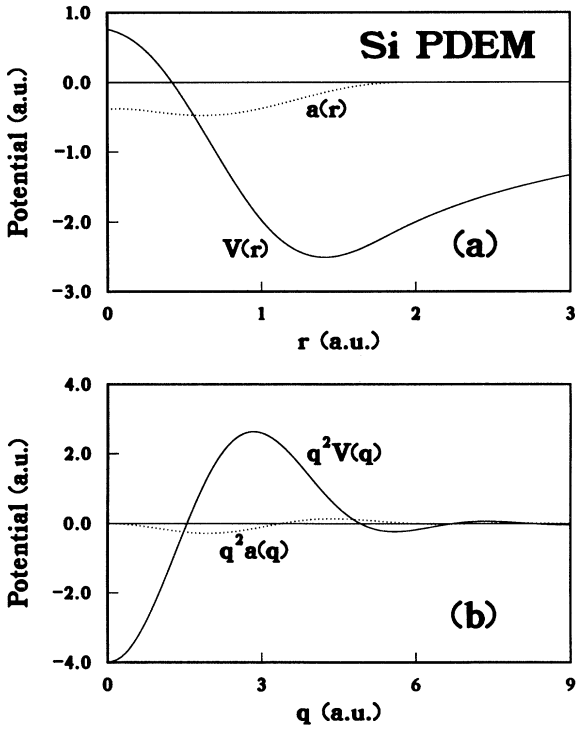

FIG. 2. The position-dependent electron-mass Si ionic pseudopotential (a) and its Fourier transform (b) obtained from the variational minimization procedure outlined in Sec. IV.

difference when compared to the more accurate nonlocal pseudopotential calculation. Reducing the eigenvalue and norm-conservation limits or the cutoff radii improved the transferability, but again resulted in a pseudopotential that was no longer computationally efficient for medium-sized cells (~ 50 atoms).

V. SEPARABLE PSEUDOPOTENTIALS

It is well known that substantial savings in computational requirements can be gained if the nonlocal potential is separable,[12] meaning that the nonlocal potential $V_{NL}^{l}(\mathbf{q},\mathbf{q}')$, which is a function dependent on both $\mathbf{q}$ and $\mathbf{q}'$, can be written as a product of a function of $\mathbf{q}$ by a function of $\mathbf{q}'$,

$$V_{NL}^{\text{sep},l}(\mathbf{q},\mathbf{q}')=\sum_{i=1}^{n_{\text{int},l}} a_i W_i(q) W_i^*(q') \sum_{m=-l}^{l} Y_{lm}(\hat{\mathbf{q}}) Y_{lm}^*(\hat{\mathbf{q}}') \ . \tag{19}$$

The savings in memory storage is made possible by no longer having to store the entire $O(n_{\text{at.}}^2 n'^2)$ nonlocal matrix, but only the $O(n_{\text{at.}}^2 n' n_{\text{int},l})$ vector elements of the functions $W_i(\mathbf{q})$. Computational operations needed for the multiplication of each nonlocal potential by the wave vector are also reduced from $O(n_{\text{at.}}^2 n'^2)$ to $O(n_{\text{at.}}^2 n' n_{\text{int},l})$. This gives substantial savings when $n_{\text{int},l} \ll n'$, but still has a quadratic dependence on $n_{\text{at.}}$. Two methods which transform any standard nonlocal pseudopotential[6,23–27] to separable form have been introduced by Štich *et al.*[13] and Kleinman and Bylander.[12]

Štich *et al.* achieves the separation by approximating the integration of the nonlocal potential in Eq. (4) with a numerical Gauss-Hermite integration formula,[13]

$$\int_0^\infty V_{NL,l}(r) j_l(|qr|) j_l(|q'r|) r^2 dr \approx \sum_i^{n_{\text{int},l}} w_i V_{NL,l}(r_i) j_l(qr_i) j_l(q'r_i) \ . \tag{20}$$

In this case the total number of operations required is reduced down to $O(n_{\text{at.}}^2 n' m_{\text{sep}})$, where

$$m_{\text{sep}}=\sum_l (2l+1) n_{\text{int},l}$$

and therefore the number of integration points should be the least number possible, in particular for the higher angular momentum components. In order to obtain an accurate answer with a small number of integration points using Gauss-Hermite, the integrand must reasonably approximate Gaussian behavior. We can avoid this restriction by using the general formulation of Gaussian integration theory[17] if $V_{NL,l}(r)$ is positive or negative definite. In this case there is for each integer $n_{\text{int},l}$, a set of weights w_i, and radii r_i such that

$$\int_0^\infty V_{NL,l}(r) f(r) dr \simeq \sum_{i=1}^{n_{\text{int},l}} w_i f(r_i) \tag{21}$$

and that the integration formula is exact for polynomials of order less than or equal to $2n_{\text{int},l}$. We have found[28] that this integration formula is numerically very accurate when used to evaluate Eq. (4). The requirement that the nonlocal potential $V_{NL,l}(r)$ be positive or negative definite can always be satisfied by an appropriate choice of the local potential $V_{\text{local}}^{PP}(r)$.

The Kleinman and Bylander[12] (KB) approach reformats the standard nonlocal potential into a form similar to the original Phillips-Kleinman[29] pseudopotential projection operator,[6]

$$V_{NL}^{\text{KB},lm}(r)=z|f(r)\rangle\langle f(r)| \ , \tag{22}$$

where

$$\langle f(r)|=\frac{\langle \Phi_l^{PP0}(r) V_{NL,l}(r)|}{|\langle \Phi_l^{PP0}(r)|V_{NL,l}(r)|\Phi_l^{PP0}(r)\rangle|^{1/2}} \ , \tag{23}$$

$\Phi_l^{PP0}(r)$ is the atomic reference pseudo-wave-function used to calculate the pseudopotential and $z=\pm 1$ is the sign of the expectation value in the denominator of Eq. (23).

This method is quite superior to the numerical integration because $n_{\text{int},l}=1$. Problems in the appearance of "ghost states" with this separation procedure have been noted.[30] In a recent paper Kleinman and Bylander[31] have given qualitative rules to avoid the existence of these ghost states, and Gonze *et al.*[32] have provided an existence theorem about the ghost states. We followed these suggestions and furthermore we also always check the logarithmic derivatives of the atom using the KB pseudopotential. For all the crystals we have studied so far with this method, we have been able to avoid ghost states for all their constituent elements (Be, C, O, Mg, Al, Si, S, Ti, Cu, Zn, Ga, As, In, Ce).

A silicon nonlocal pseudopotential was generated using the method of Ref. 6, in the non-spin-polarized ground-state valence configuration $3s^2 3p^2 3d^0$. The cutoff radii were $r_{cs}=r_{cp}=r_{cd}=2.1a_0$; the exchange correlation was Ceperley-Alder[20] as parametrized by Perdew and Zunger.[21] The properties of silicon calculated using an energy cutoff of $E_{\text{cut}}=16$ Ry are shown in Table I. These results are virtually identical to what we obtained using the semilocal form [Eq. (4)] of the pseudopotential. The ionic pseudopotential and its Fourier transform are shown in Figs. 3(a) and 3(b), respectively. The Gonze *et al.* theorem[32] revealed that no ghost states would exist if the local potential was chosen as the *s*, *p*, or *d* component. In order to reduce the number of angular components, the highest angular component is usually taken as the local potential. However, in this case we selected the *p* component as the local as it produced the best overall transferability for the total pseudopotential as measured by its logarithmic energy derivatives. In Fig. 4(a) we show the normalized right side of the KB operator [Eq. (23)] and in Fig. 4(b), its reciprocal space counterpart.

While these methods do greatly reduce the amount of needed storage and operations, they will scale quadratically with $n_{\text{at.}}$. Two papers have appeared recently[10,11] that discuss the efficient calculation of the matrix elements for a semilocal pseudopotential [Eq. (4)]. Both methods use FFT's so that the calculation of $\hat{V}_{NL}\Psi$ is performed in real space. The method of Goedecker and

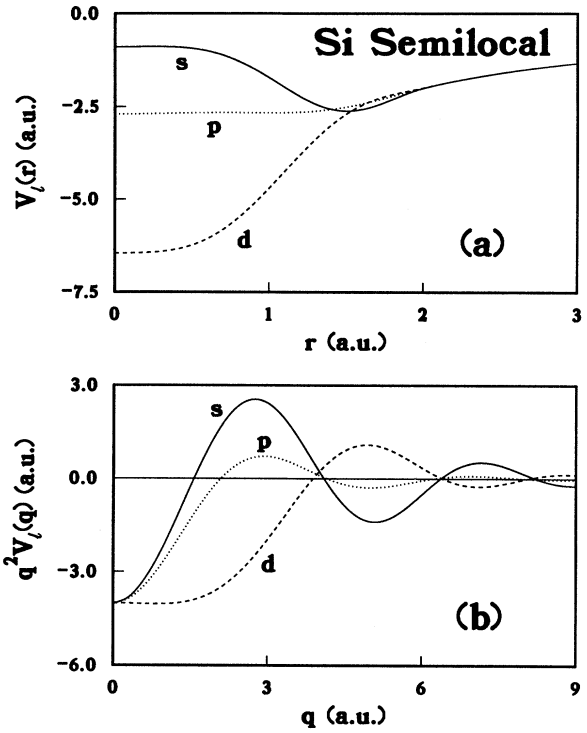

FIG. 3. The semilocal pseudopotential (a) and its Fourier transform (b) used in Sec. V. The pseudopotential was obtained using the method of Ref. 6.

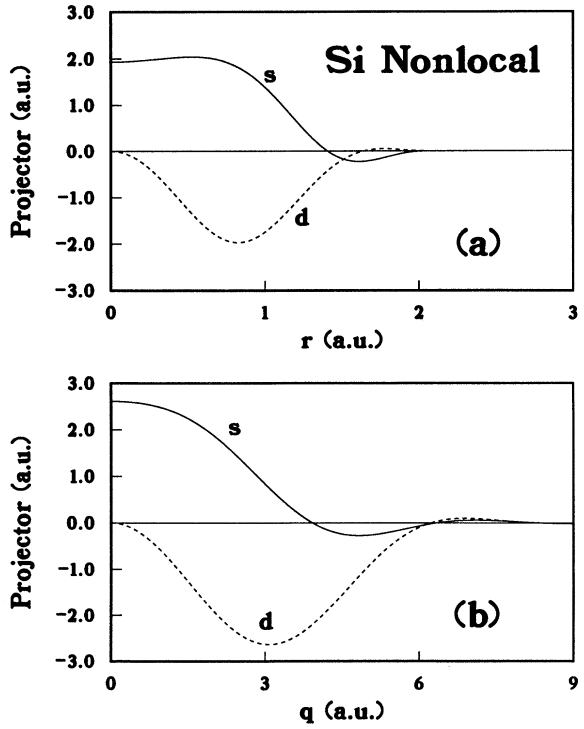

FIG. 4. The real- (a) and reciprocal-space (b) normalized nonlocal Kleinman and Bylander (Ref. 12) projection operators [Eq. (23)] for the silicon semilocal pseudopotential shown in Fig. 3.

Maschke[11] scales with $n_{\text{at.}}^{4/3}$, but only for high-symmetry **k** points such as the Brillouin zone center Γ. Since the simulation of liquids and amorphous systems with large unit cells used within the Car and Parrinello method[2] uses only the Γ point, this is not as restrictive as it may first appear. The method of Gonze *et al.*[10] uses a polynomial interpolation of the wave function in the FFT grid mesh around each atom to calculate the product $\hat{V}_{NL}\Psi$ and its computational load scales with $n_{\text{at.}}\ln n_{\text{at.}}$. We were unable to show that the numerical interpolation of the wave function conserves the Hermiticity of the pseudopotential operator; this last property is important for the stability of some iterative diagonalization methods.[33]

The most efficient way to calculate the product $\hat{V}_{NL}\Psi$ is obtained if we use the concept of a dual representation of the wave function in both real and reciprocal space.[1] The main idea is that the representation of the wave function by its Fourier components at a finite number of reciprocal-lattice vectors, $\mathbf{G}_i$, is equivalent to giving the value of the wave function on a discrete uniform three-dimensional grid of points $\mathbf{s}_i$ in the primitive cell. The change in representation using fast Fourier transforms takes only $O(N'\ln N')$ operations, where N' is the number of grid points, which is only slightly larger than the basis-set size. In the real-space representation the nonlocal potential is given by a matrix where each row and column is associated with a grid point in the unit cell. If the nonlocal potential has a finite range such that the nonlocal potentials from different atoms in the unit cell do not overlap, then one can expect that the matrix elements would be nonzero if and only if the grid points corresponding to the row and column indices are both within the finite range of the interaction for that particular atom. This means that by a suitable renumbering of the rows and columns the nonlocal potential matrix is block diagonal with all the nonzero elements concentrated on $n_{\text{at.}}$ square blocks along the diagonal,

$$\sum_{n_{\text{at.}}}\sum_{l} V_{NL}^{l}(\mathbf{r},\mathbf{r}')=\begin{bmatrix} B_1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & & & \vdots \\ 0 & 0 & B_2 & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & B_{n_{\text{at.}}} & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 \end{bmatrix}. \tag{24}$$

The size of each block, B_i, depends only on the number of grid points around each atom. This number depends on the energy cutoff E_{cut} but not on the number of atoms in the unit cell. The product $\hat{V}_{NL}\Psi$ can therefore be calculated with $O(n_{\text{at.}}n'^2)$ operations in real space, and therefore has a linear scaling in the number of atoms. For a separable pseudopotential the matrix blocks can be written as a sum of projectors and we would have a better scaling for large energy cutoffs, since the number of operations in that case would scale as $O(n_{\text{at.}}n')$.

For the sake of notational simplicity, let us consider only the effect of one of the KB projectors from one of the atoms in the unit cell on a trial wave function. This

operator can be written as

$$V_{NL}^{l}(\mathbf{r},\mathbf{r}')=\sum_{\mathbf{t}} zf(\mathbf{r}-\mathbf{t})f^{*}(\mathbf{r}'-\mathbf{t}) , \tag{25}$$

where the functions $f(\mathbf{r})$ and z are defined in Eq. (23), and the sum is over all the lattice sites of the crystal. It can then be shown that

$$\begin{aligned}\hat{V}_{LN}\Psi(\mathbf{r})&=\int_{\text{crystal}} V_{NL}^{l}(\mathbf{r},\mathbf{r}')\Psi_{\mathbf{k}}(\mathbf{r}')d\mathbf{r}' \\ &=ze^{i\mathbf{k}\cdot\mathbf{r}}g(\mathbf{r})\int_{\text{cell}} g^{*}(\mathbf{r}')u(\mathbf{r}')d\mathbf{r}' \\ &=e^{i\mathbf{k}\cdot\mathbf{r}}v(\mathbf{r}) ,\end{aligned} \tag{26}$$

where the last equality defines $v(r)$,

$$g(\mathbf{r})=\sum_{\mathbf{t}} f(\mathbf{r}-\mathbf{t})e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{t})} , \tag{27}$$

$u(\mathbf{r})$ is defined by $\Psi_{\mathbf{k}}(\mathbf{r})=e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r})$, and $\mathbf{t}$ are the lattice translation vectors. Note that $u(\mathbf{r})$, $g(\mathbf{r})$, and $v(\mathbf{r})$ all have the crystal periodicity. The function under the integral in Eq. (26) is periodic and can therefore be efficiently integrated numerically in the uniform FFT grid, that is, using Fourier's integration.[17] If $\mathbf{s}_i$ denotes one of these grid points, then we have

$$v(\mathbf{s}_i)\simeq zg(\mathbf{s}_i)\frac{1}{N'}\sum_{j=1}^{N'} g^{*}(\mathbf{s}_j)u(\mathbf{s}_j) . \tag{28}$$

Because $f(\mathbf{r})$ has a finite range, the number of grid points where $g(\mathbf{s}_j)\neq 0$ is $O(n')$ and is independent of the size of the unit cell. An important property of this numerical approximation [Eq. (28)] for the nonlocal operator is that it is Hermitian and therefore does not introduce unwanted instabilities in iterative diagonalization procedures.[33]

VI. CONCLUSIONS

In this study we have investigated the computational dependence on the number of atoms per unit cell for local, position-dependent electron-mass (PDEM) and separable nonlocal pseudopotentials. We used a variational minimization procedure to generate the local and PDEM pseudopotentials. These pseudopotentials were generated such that they satisfied eigenvalue and norm-conservation constraints within a preset error tolerance. However, in order to make the local and PDEM pseudopotentials "soft" enough to be more efficient than the separated nonlocal pseudopotential methods for medium-size unit cells (less than 50 atoms), a compromise in the pseudopotential transferability had to be made. While general chemical trends could still be determined with these pseudopotentials, the determination of the energy difference between structures was questionable. The calculation in real space of $\hat{V}_{NL}\Psi$ is the most promising of the three nonlocal methods we examined, because it is simple [Eqs. (25)–(28)] and does not require compromises on its transferability. This method should be particularly useful for simulations of large unit cells with more than ~ 50 atoms.

ACKNOWLEDGMENTS

We are grateful to Professor J. R. Chelikowsky for stimulating discussions and to Dr. W. M. C. Foulkes for sending us Ref. 9 prior to publication. We also benefited from discussions with Dr. S. Froyen and Dr. G. Bachelet. This work was supported by a computer time grant from the Minnesota Supercomputer Institute.

[1] J. L. Martins and M. L. Cohen, Phys. Rev. B **37**, 6134 (1988).
[2] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[3] M. P. Teter, M. C. Payne, and D. C. Allan, Phys. Rev. B **40**, 12 255 (1989).
[4] M. C. Payne, J. D. Joannopoulos, D. C. Allan, M. P. Teter, and D. H. Vanderbilt, Phys. Rev. Lett. **56**, 2656 (1986).
[5] Here we indicate by $O(N)$ a quantity that is proportional to N with a coefficient of proportionality of the order of unity. This coefficient would depend on the specific details of the calculation and architecture of the computer. For example, a complex matrix requires twice the storage of a real matrix; a symmetric matrix may be stored in "packed" form with a 50% memory storage savings or stored in the "traditional" form for efficient vectorization.
[6] N. Troullier and J. L. Martins, Solid State Commun. **74**, 613 (1990); Phys. Rev. B **43**, 1993 (1991).
[7] Th. Starkloff and J. D. Joannopoulos, Phys. Rev. B **19**, 1077 (1979).
[8] G. B. Bachelet, D. M. Ceperley, and M. G. B. Chiocchetti, Phys. Rev. Lett. **62**, 2088 (1989).
[9] W. M. C. Foulkes and M. Schlüter, Phys. Rev. B **42**, 11 505 (1990).
[10] X. Gonze, J. P. Vigneron, and J.-P. Michenaud, J. Phys. Condens. Matter **1**, 525 (1989).
[11] S. Goedecker and K. Maschke, Phys. Rev. B **41**, 3230 (1990).
[12] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).
[13] I. Štich, R. Car, M. Parrinello, and S. Baroni, Phys. Rev. B **39**, 4997 (1989).
[14] J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).
[15] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).
[16] Following tradition, we will use the same symbol for a function and its Fourier transform. We also use the same symbol for an operator and its representation. Arguments or subscripts and context usually suffice to distinguish between them.
[17] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University, Cambridge, England, 1986).
[18] Full orthogonalization of the eigensolutions will still require $O(M^2N)$ operations. Although $M^2N \ll MN^2$, this number still scales with the cube of the number of atoms, and therefore efficient iterative diagonalization algorithms should minimize the number of orthogonalization steps.
[19] Equation (3.5) in N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[20] D. M. Ceperley and B. J. Adler, Phys. Rev. Lett. **45**, 566 (1980).
[21] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[22] F. D. Murnaghan, Proc. Natl. Acad. Sci. U.S.A. **30**, 244

(1944).
[23]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).
[24]G. P. Kerker, J. Phys. C **13**, L189 (1980).
[25]A. Zunger and M. L. Cohen, Phys. Rev. B **18**, 5449 (1978).
[26]G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).
[27]D. Vanderbilt, Phys. Rev. B **32**, 8412 (1985).
[28]J. L. Martins, Phys. Rev. B **41**, 7883 (1990); **38**, 12 776 (1988).
[29]J. C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959).
[30]D. R. Hamann, Phys. Rev. B **40**, 2980 (1989).
[31]D. M. Bylander and L. Kleinman, Phys. Rev. B **41**, 907 (1990).
[32]X. Gonze, P. Käckell, and M. Scheffler, Phys. Rev. B **41**, 12 264 (1990).
[33]H. Rutishauser, Numer. Math. **13**, 4 (1969).