Accuracy of the chemical-pseudopotential method for tetrahedral semiconductors

W. M. C. Foulkes
Blackett Laboratory, Imperial College of Science, Technology and Medicine, Prince Consort Road, London SW7 2BZ, United Kingdom
(Received 22 June 1993; revised manuscript received 5 August 1993)

Simple two-center tight-binding models have often been used in electronic-structure calculations, although uncertainty about the foundations of the tight-binding method has severely limited progress. This uncertainty has been much reduced following the recent successes of the non-self-consistent Harris scheme but several important questions still remain. In particular, the accuracy of the (almost ubiquitous) two-center approximation is still in doubt. It has often been argued that although this is a poor approximation when applied crudely, it can be justified within the more sophisticated context of chemical-pseudopotential theory. The argument looks sensible, but clear quantitative tests only became possible with the advent of the Harris scheme. This paper reports a careful series of tests of the accuracy of the two-center approximation for germanium, and shows that neither the crude version nor the more sophisticated chemical-pseudopotential version is accurate enough to be useful. The remarkable success of two-center tight binding in predicting the structures and surface reconstructions of semiconductors is, therefore, still mysterious.

I. INTRODUCTION

If electronic structure theory is ever to be of more than marginal interest to materials scientists, it will be necessary to develop reliable and trustworthy methods to calculate the forces between the atoms in very complicated solids. We can already calculate the physical properties of the majority of perfect crystals, but this is only a beginning and materials scientists are waiting for the day when it will be possible to use quantum mechanics to calculate all the forces between the atoms around a moving dislocation. That day may come quite soon for solids made up of atoms with reasonably weak pseudopotentials,[1,2] but for other materials, among them the transition metals, the prospects are not so good. The usual approach is to resort to classical interatomic pair potentials, but although these often give qualitatively useful results they are unlikely to be quantitatively accurate in most solids.

What is required is an armory of methods to fill the gap between full self-consistent density functional theory (DFT) and pair potentials, and several such methods are already being developed. Among these are a number of schemes for generating classical many-atom potentials with forms suggested by quantum mechanics.[3–10] This approach has already demonstrated its worth, and although it may fail in some systems — the classical potentials required may be so complicated that they are unusable — there have already been many useful applications.

At the other end of the scale, the recent successes of non-self-consistent density functional (DF) calculations based on the Harris functional[11–16] have suggested that it is often possible to simplify DF methods considerably without losing much accuracy. A number of methods[17–19] have been developed which are reasonably accurate and fully quantum mechanical (hence they still require the solution of a matrix eigenvalue problem), but which are a good deal easier than self-consistent DF calculations. Their successes may point the way to even simpler schemes, and perhaps at long last allow the development of reliable tight-binding (TB) models. This paper discusses some of the remaining problems which must be overcome before such a goal can be realized.

The semiempirical TB method[20,21] has been one of the most widely used electronic-structure methods in solids and has been particularly valuable in complicated systems for which more sophisticated calculations are difficult. Although usually applied to transition metals, TB models have also been used to predict surface reconstructions[22] and grain boundary structures[23] in semiconductors, and have been shown capable[24] of reproducing the whole "phase diagram" (the energy versus volume curves for all plausible low energy crystal structures) of silicon. TB calculations are comparatively straightforward and yield a simple and appealing picture of bonding in terms of orbital occupations and bond orders. In addition, they often work better for "difficult" solids such as transition metals than for "easy" solids such as Al. These advantages are partially offset by the fact that TB models are often inaccurate and sometimes untrustworthy, but the method has an important place in the electronic-structure calculator's armory.

This work follows on from an earlier paper[12] in which we showed how the Harris functional may be used to derive simplified but accurate electronic-structure methods which look very like semiempirical TB models. The ideas in that paper underpin the electronic-structure methods developed by Gibson and Haydock,[18] and (independently) by Sankey[17] and by Harris and Hohl.[19] These have no adjustable parameters and have been very successful, but differ from most semiempirical methods in two important ways: first, they use a basis set of nonorthogonal (atomiclike) orbitals and so the calculations

0163-1829/93/48(19)/14216(10)/$06.00

involve an overlap matrix as well as a Hamiltonian matrix; and second, the Hamiltonian matrix elements do not have a simple two-center form. Instead, they involve complicated three-center integrals which depend on the relative positions of three different atoms.

Since semiempirical TB is often very successful, it is tempting to ask whether it is really necessary to include three-center terms and an overlap matrix? The question is an old one, and the general consensus[20] is that (at least for tetrahedral semiconductors) neither the overlap matrix nor the three-center terms can be ignored. So why was Goodwin's[24] two-center orthogonal TB model for Si so successful? The usual answer is that the matrix which looks like the Hamiltonian in his model is actually a parametrization of the D matrix of chemical-pseudopotential (CP) theory.[25–28,20] If this is correct, then it should be possible to derive simple two-center D matrix TB models from first principles and so obtain an electronic-structure method with all the virtues of semiempirical TB but without its drawbacks. The scheme would have no fitting parameters, would be trustworthy when the Harris functional is trustworthy (which is surprisingly often), and all the approximations would be clear and quantifiable.

Following the development of the Harris functional, it is now possible to isolate the main approximation (the assumption that the three-center contributions to the D matrix are negligible) inherent in the CP approach and see how well it stands up to quantitative scrutiny. This is the aim of this paper, and the result is that the approximation is not accurate enough to be useful in tetrahedral semiconductors: if anything, it is worse than simply ignoring all the three-center contributions to the ordinary Hamiltonian matrix. This does not imply that the method fails for all materials — and indeed the work of Bullett[20] has shown that it often works well for the d bands of transition metals and in other materials commonly thought of as being in the TB (narrow band) limit — but the arguments cannot be extended to the nearly-free-electron solids usually treated with plane waves. The conventional justification for using two-center TB in such materials is no good, and the more sophisticated D matrix version is no better than the simple Hamiltonian matrix version.

This paper contains seven sections altogether. After this introduction comes a background section which describes semiempirical TB and the Harris functional and clarifies the relationship between the two. This discussion contains little that is new and may be skipped by readers already familiar with the material; however, it is necessary to define the framework and notation on which the rest of the work is based and should help to make the paper more comprehensible to readers who are not experts in the field. Section III describes the results of some plane wave and pseudopotential calculations which were done to test the accuracies of both the Harris functional and the local Ge pseudopotential used throughout this work. Section IV investigates the crude (Hamiltonian matrix) two-center approximation for a number of different basis sets, and finds, not surprisingly, that it is not very accurate. Section V explains the CP approach and discusses how it may be used to justify two-center TB. Section VI gives the results of using the two-center D matrix approximation for Ge, and some conclusions are drawn in Sec. VII.

II. BACKGROUND

A. Semiempirical tight-binding models

Semiempirical TB (Refs. 29 and 20) predates DFT by several years and was one of the first widely used methods for calculating solid state band structures and interatomic forces. It is assumed that the total energy may be written in the form

$$E = \sum_{i=1}^{N} \varepsilon_i + \frac{1}{2} \sum_{\alpha} \sum_{\beta \neq \alpha} U(|\mathbf{R}_\alpha - \mathbf{R}_\beta|) \quad , \qquad (1)$$

where N is the total number of electrons in the system, $\mathbf{R}_\alpha$ is the position of an atom, and U is a pair potential representing the repulsive interactions between the ionic cores. The eigenvalues ε_i are supposed to be the solutions of a non-self-consistent independent electron Schrödinger equation but this is never solved directly. Instead, one imagines solving it variationally using a basis set of localized atomiclike orbitals. The matrix elements of the corresponding Hamiltonian (H) and overlap (S) matrices are treated as free parameters and are adjusted to reproduce the results of experiments or other calculations. For simplicity, the S matrix is often taken to be the identity (the basis functions are assumed to be orthonormal Wannier functions) and the H matrix elements are taken to be zero beyond nearest or next nearest neighbor distances.

These few approximations are enough to allow calculation of the eigenvalues and hence the band structure, but if interatomic forces are required, then one also has to know how the matrix elements and pair potentials depend on the atomic positions. Most models assume that the pair potential and all matrix elements linking the atoms at $\mathbf{R}_\alpha$ and $\mathbf{R}_\beta$ depend only on the distance between those two atoms and are independent of the positions of other atoms in the vicinity (this is the two-center approximation). The distance dependence is usually taken to be a simple power law or an exponential with one or two adjustable parameters to be determined empirically. Furthermore, the possible environment dependence of the intra-atomic matrix elements is often ignored (neglect of crystal field terms). These approximations are not easy to justify, but are necessary to keep the number of adjustable parameters down and have been very successful in practice.

Any attempt to derive nonempirical parameter-free TB models will have to reconsider the validity of all the above assumptions and approximations, and must therefore start from some more fundamental theory. The obvious starting point is DFT within the local density approximation,[30,31] and the Harris functional[11,12] turns out to provide the clues needed.

B. The Harris functional

In the original development of DFT, Hohenberg and Kohn[30] and Kohn and Sham[31] showed that the total ground state energy of a set of interacting electrons moving in an external potential, $\hat{V}_{\rm ext}$ (usually the Coulomb potential due to the nuclei or inert atomic cores), is the minimum value of a functional, $E[n]$, of the electron number density, $n(\mathbf{r})$. To evaluate this functional at a given density, $n_{\rm out}$, one first has to find the effective one-electron potential, $\hat{V}_{\rm in}$, which *generates* $n_{\rm out}$. This means that $n_{\rm out}$ must be the density obtained by solving a set of Schrödinger-like independent electron equations known as the Kohn-Sham equations (Hartree atomic units are used unless otherwise indicated):

$$\left(-\frac{1}{2}\nabla^2 + V_{\rm in}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}),$$

$$n_{\rm out}(\mathbf{r}) = \sum_{i=1}^{N}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}) \quad . \tag{2}$$

For simplicity, one usually reverses the problem and works by choosing $\hat{V}_{\rm in}$ and then calculating $n_{\rm out}$ from Eq. (2). Once $n_{\rm out}$ has been found, the total energy functional can be evaluated using

$$E[n_{\rm out}] = \sum_{i=1}^{N}\varepsilon_i - \int[V_{\rm in}(\mathbf{r}) - V_{\rm ext}(\mathbf{r})]n_{\rm out}(\mathbf{r})d^3r + \frac{1}{2}\int\int\frac{n_{\rm out}(\mathbf{r})n_{\rm out}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r\,d^3r' + E_{\rm xc}[n_{\rm out}] \ , \tag{3}$$

where $E_{\rm xc}[n]$ is the exchange-correlation functional which is not known exactly (all the complicated aspects of the many-body problem are hidden in it) but for which simple local density approximations (LDA's) work well in most solids.[32]

It can be shown that $E[n]$ attains its minimum value when $\hat{V}_{\rm in}$ and $n_{\rm out}$ satisfy a "self-consistency" condition of the form $\hat{V}_{\rm in} = \hat{V}_{\rm KS}[n_{\rm out}]$, where

$$V_{\rm KS}([n],\mathbf{r}) = V_{\rm ext}(\mathbf{r}) + \int\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r' + \mu_{\rm xc}([n],\mathbf{r}) \tag{4}$$

is a density-dependent effective potential known as the Kohn-Sham potential and $\mu_{\rm xc}([n],\mathbf{r})$ is the first functional derivative of $E_{\rm xc}[n]$. Ground state properties may therefore be calculated by an iterative procedure in which the input potential is adjusted until self-consistency is attained.

Since the Hohenberg-Kohn energy functional is minimized at the ground state density n_0, a variational principle applies: if $n_{\rm out}$ is a trial ground state density, then $E[n_{\rm out}]$ is greater than or equal to $E[n_0]$ and the error is second order in $n_{\rm out} - n_0$. A good approximate ground state density therefore leads to an even better approximate ground state energy and this can be very helpful in non-self-consistent calculations. Unfortunately, the expression for $E[n_{\rm out}]$ involves Hartree and exchange-correlation terms evaluated at $n_{\rm out}$ and constructing these potentials is not easy when using basis functions resembling atomic orbitals or Wannier functions. With the small basis sets considered here, solving the matrix eigenvalue problem is not necessarily much harder than constructing the potentials (although this depends on the size of the system) and so one might as well iterate to self-consistency.

If this was the end of the story, then non-self-consistent DFT would be of limited use in this work. However, it was realized initially by Harris[11] and later independently by Foulkes and Haydock[12] that a further simplification can be made. A sensible way of choosing the input potential is to make a guess, $n_{\rm in}$, at the ground state density, and then to set $\hat{V}_{\rm in} = \hat{V}_{\rm KS}[n_{\rm in}]$. If this is done and the terms in Eq. (3) which depend explicitly on $n_{\rm out}$ are expanded in the small quantity $n_{\rm out} - n_{\rm in}$, then it can be shown that the alternative energy functional (now usually known as the Harris functional),

$$E_H[n_{\rm in}] = \sum_{i=1}^{N}\varepsilon_i - \frac{1}{2}\int\int\frac{n_{\rm in}(\mathbf{r})n_{\rm in}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r\,d^3r' + E_{\rm xc}[n_{\rm in}] - \int\mu_{\rm xc}([n_{\rm in}],\mathbf{r})n_{\rm in}(\mathbf{r})d^3r \ , \tag{5}$$

differs from $E[n_{\rm out}]$ only by terms which are of second (or higher) order in $n_{\rm in} - n_{\rm out}$. The Harris functional is therefore also stationary about the ground state density and so there is no reason to expect that it should be less accurate than the Hohenberg-Kohn functional when used in non-self-consistent calculations. Its main advantage is that the output density does not appear explicitly.

Perhaps surprisingly, it turns out that the Harris functional approximation to the ground state energy is usually better than the variational estimate obtained using the Hohenberg-Kohn functional with the same input potential, and the accuracy obtainable is often remarkable.[11,12] Using simple input densities constructed by superposing neutral atoms, lattice parameters, elastic moduli, phonon spectra, and "phase diagrams" have been calculated for a wide variety of solids with errors not much larger than those due to the LDA.[13–16] There have been occasional failings (at surfaces, for example, when it can be hard to guess sensible trial input densities) but superposing neutral atomic densities works well enough for many purposes.

C. Tight-binding models and the Harris functional

The Harris functional energy is the sum of three contributions: the eigenvalue sum, the "double-counting" energy [the other terms in Eq. (5)], and the Coulomb replusion energy of the nuclei (which has not been considered until now). Both the one-electron Hamiltonian and the double-counting energy are particularly simple if the input density takes the form

$$n_{\rm in}(\mathbf{r}) = \sum_{\alpha} n_\alpha(\mathbf{r}) \quad , \tag{6}$$

where n_α is a spherical atomlike electron density on the atom at $\mathbf{R}_\alpha$ and contains exactly Z_α electrons (Z_α is the nuclear charge). The work cited in Sec. II B shows that choosing the input density this way works well in many cases.

The sum of the double-counting and Coulomb repulsion energies may then be written in the form

$$\sum_\alpha U_\alpha + \frac{1}{2}\sum_\alpha \sum_{\beta(\neq\alpha)} U_{\alpha\beta}(|\mathbf{R}_\alpha - \mathbf{R}_\beta|) + U_{\rm np} \quad , \qquad (7)$$

where U_α is a constant intra-atomic contribution and $U_{\alpha\beta}$ is a pair potential which vanishes rapidly as $|\mathbf{R}_\alpha - \mathbf{R}_\beta|$ tends to infinity. U_α and $U_{\alpha\beta}$ can be evaluated from monomer and dimer calculations, respectively, and stored for later use in the solid. $U_{\rm np}$ is a nonpairwise contribution containing only exchange-correlation terms and is more difficult to precalculate and store, although it may be important in some cases.[33] In the LDA it arises because the exchange-correlation potential at $\mathbf{r}$ is a nonlinear function of the electron density there.

The Harris functional prescription for the form of the independent particle Hamiltonian is equally simple: it is just the usual Kohn-Sham Hamiltonian evaluated at the input density and may be written in the form

$$\hat{H} = \hat{T} + \sum_\alpha \hat{V}_\alpha + \hat{W} \quad , \qquad (8)$$

where $\hat{T}$ is the kinetic energy operator, $\hat{V}_\alpha$ is the Kohn-Sham potential (including Hartree and exchange-correlation terms) for an isolated atom at $\mathbf{R}_\alpha$, and $\hat{W}$ arises from the nonlinearity of the exchange-correlation terms,

$$W(\mathbf{r}) = \mu_{\rm xc}\left(\left[\sum_\alpha n_\alpha\right], \mathbf{r}\right) - \sum_\alpha \mu_{\rm xc}([n_\alpha], \mathbf{r}) \quad . \qquad (9)$$

Most of the fundamental questions about TB have now been answered: the absence of self-consistency has been explained, the replulsive pair potential has been identified, and the form of the independent electron Hamiltonian is known. The remaining questions concern the choice of basis set and the evaluation of the H and S matrix elements. The work of Sankey,[17] Gibson, Haydock, and LaFemina,[18] and Harris and Hohl[19] suggests that a basis of atomic (or atomiclike) orbitals, $\phi_{\alpha i}$, is often adequate (α labels the atom and i specifies the particular basis function on that atom). The S matrix elements, $S_{\alpha i,\beta j} = \langle\phi_{\alpha i}|\phi_{\beta j}\rangle$, are then environment independent (intra-atomic elements being constants and interatomic elements depending only on the distance between the two atoms involved), although the H matrix elements, $H_{\alpha i,\beta j} = \langle\phi_{\alpha i}|\hat{H}|\phi_{\beta j}\rangle$, are not. Ignoring $\hat{W}$ for the time being, the intra-atomic H matrix elements involve one-center terms, $\langle\phi_{\alpha i}|\hat{T} + \hat{V}_\alpha|\phi_{\alpha j}\rangle$, and two-center crystal field terms which depend on the positions of neighboring atoms, $\langle\phi_{\alpha i}|\hat{V}_\beta|\phi_{\alpha j}\rangle$. The interatomic matrix elements include two-center terms, $\langle\phi_{\alpha i}|\hat{T} + \hat{V}_\alpha + \hat{V}_\beta|\phi_{\beta j}\rangle$, and three-center terms, $\langle\phi_{\alpha i}|\hat{V}_\gamma|\phi_{\beta j}\rangle$. All overlap, crystal field, and one- and two-center H matrix elements can be obtained from monomer and dimer calculations, tabulated as functions of the interatomic distance, and looked up whenever needed. The three-center terms require a series of trimer calculations and the tabulation of a number of more complicated two-dimensional functions.[17]

The TB scheme described in this section is essentially the one used in Refs. 17, 18, and 19, and differs in several important respects from the simple semiempirical TB models described earlier. The main differences are that semiempirical schemes usually assume an orthogonal basis set and ignore all three-center and crystal field terms. While including crystal field terms and an overlap matrix complicates the method only slightly, including the three-center integrals is more unpleasant.

III. TESTS OF THE HARRIS FUNCTIONAL FOR GERMANIUM

This paper will concentrate on Ge, both because it is a tetrahedral semiconductor for which one might expect semiempirical TB to work quite well, and because the Ge ionic cores can be well represented using local pseudopotentials (making it possible to avoid a number of uninteresting technical issues). The local pseudopotential used here is of the Starkloff-Joannopoulos[34] form,

$$V_{ps}(r) = -\frac{Z_v}{r}\left(\frac{1 - e^{-\lambda r}}{1 + e^{\lambda(r_c - r)}}\right) \quad , \qquad (10)$$

with $Z_v = 4$, $\lambda = 18a_0^{-1}$, and $r_c = 1.0a_0$.

Tests of the accuracy of the pseudopotential and the Harris functional were carried out using a conventional plane wave based total energy program. All calculations used 10 special k points in the irreducible wedge of the Brillouin zone and the Ceperley-Alder[35,36] form of the LDA. Kinetic energy cutoffs up to 12.5 Hartrees were investigated and the calculations were well converged. The total energy in the diamond structure was calculated at a range of different lattice parameters and a Murnaghan equation of state[37] was fitted to the results. The equilibrium lattice parameter and bulk modulus were 5.53 Å and 780 kbar, respectively, agreeing quite well with the experimental values[38] of 5.65 Å and 770 kbar.

As well as these self-consistent calculations, other calculations were done to compare the accuracies of the Hohenberg-Kohn and Harris functionals when used non-self-consistently. The input density chosen was a superposition of pseudoatomic valence densities, the Kohn-Sham equations were solved once only, and the corresponding Hohenberg-Kohn and Harris energies, $E[n_{\rm out}]$ and $E_H[n_{\rm in}]$ [see Eqs. (3) and (5), respectively], were evaluated. The self-consistent and non-self-consistent results are compared in Table I. As expected, the non-self-consistent Hohenberg-Kohn energy is a little higher than the self-consistent energy, but both the energy and the bulk modulus are very close to the self-consistent values. The Harris functional total energy is lower than the self-consistent value (showing that the Harris functional is certainly not minimized at self-consistency in the LDA) and the bulk modulus is about 5% higher. The Hohenberg-Kohn functional performs slightly bet-

TABLE I. The (pseudopotential) total energy, lattice parameter, and bulk modulus of Ge in the diamond structure. The results of self-consistent (SC) and non-self-consistent (NSC) calculations using the Hohenberg-Kohn (HK) and Harris (H) functionals are shown.

	Energy (Hartrees)	a (Å)	B (kbar)
SC	-7.891	5.53	780
NSC-HK	-7.889	5.55	770
NSC-H	-7.897	5.50	820

ter than the Harris functional in this example, but both are accurate enough to be useful. The non-self-consistent band structure, shown in Fig. 1, is almost indistiguishable from the self-consistent band structure (not shown). This close similarity is surprising since the stationary and variational principles discussed earlier refer to total energies and do not apply to eigenvalues alone.

These calculations have shown that both the local pseudopotential and the Harris functional with superposed atomic input densities work well for germanium in the diamond structure. Any large errors in the TB calculations which follow must therefore be attributable to the limited basis set used or to the two-center approximation.

IV. THE SIMPLE TWO-CENTER APPROXIMATION

In Ref. 12, it was found that using the Harris functional and a basis set of pseudoatomic s and p orbitals gave a reasonable quantitative description of the Ge dimer, which was improved by adding a simple analytic d orbital to the basis set. If Ge acts like C and Si, then the work of Sankey[17] suggests that similar basis sets will also be adequate in the solid. Assuming this to be the case, a TB model of the type discussed in Sec. II C was constructed. Since the main aim of this work was to assess the accuracy of the simple two-center approximation, all three-center contributions to the matrix elements and the double-counting energy were neglected. All one- and two-center terms (including the two-center parts of the matrix elements of $\hat{W}$ — see Sec. II C) were obtained from monomer and dimer calculations.

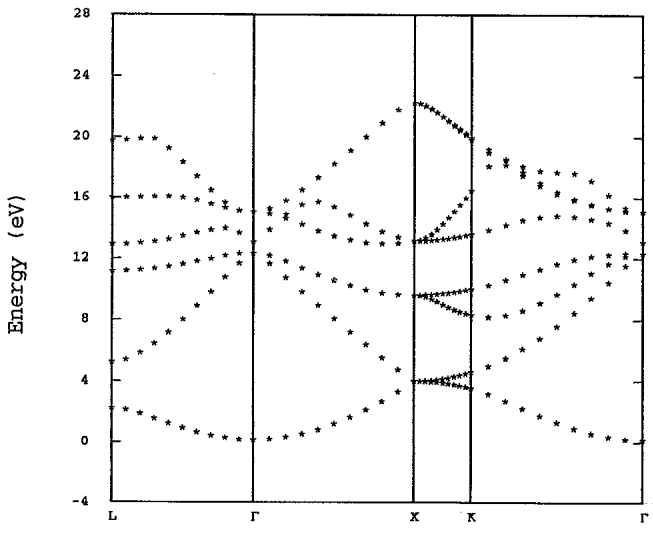

FIG. 1. Non-self-consistent Ge band structure. The input potential was constructed by superposing free atomic electron densities.

The first basis set tried consisted of the single atomic $4s$ and three atomic $4p$ orbitals on each atom ($1s$ and $2p$ orbitals of the pseudoatom). Calculations for Ge dimers[12] (when there are no three-center terms to worry about) gave an interatomic spacing of $4.9a_0$ and a vibrational stretch frequency of 240 cm^{-1}, to be compared to the accurate DF results[39] of $4.75a_0$ and 240 cm^{-1}. In the solid, however, this basis set yields matrix elements which extend to unreasonable distances. Even when all matrix elements between atoms less than 9.53 Å apart are included (each atom then interacts with 158 neighbors—to be compared to the 4 nearest neighbors included in most semiempirical calculations), the band structure obtained bears little if any relation to the accurate Harris functional band structure. Adding d orbitals to the basis set reduces the dimer interatomic spacing to $4.73a_0$ but does not improve the band structure noticeably.

What is the cause of the large band-structure errors? Some of the difficulties could be due to the two-center approximation but experiment shows that there is a more immediate problem: increasing the cutoff a little makes a big difference to the results. Similar observations of sensitive dependence on the inclusion of matrix elements to very distant neighbors have been made by Sankey[17] but the phenomenon is surprising nevertheless. The reason for the sensitivity, it turns out, is that the overlap matrix is very nearly singular. This means that it is possible to construct states $|\chi\rangle = b_{\alpha i}|\phi_{\alpha i}\rangle$ (repeated suffixes are to be summed) which have large expansion coefficients, $b_{\alpha i}$, but almost zero norm. One consequence of this near linear dependence of the basis functions is that the eigenvalues can become very sensitive to small changes in the H matrix, exactly as observed.

It would be interesting to investigate whether this linear dependence problem is restricted to tetrahedral semiconductors or whether it always arises when trying to use atomic orbital basis sets in nearly-free-electron solids (more naturally described using plane waves). In principle, of course, the entire set of (bound and unbound) atomic orbitals on just one atom makes up a complete basis for the whole solid, and so it is plausible to suppose that using a number of atomic orbitals on each site often leads to near linear dependence. This could explain why such methods have never really taken off in solid state calculations, and suggests that it may be worth working out how to generate localized basis sets which, like plane waves, can be increased to span the entire Hilbert space, but which never become overcomplete. Then one would have a really well founded numerical method in which the number of basis functions could be systematically increased until convergence was obtained without encountering difficult linear dependence problems.

Some of these problems could probably be overcome by

using the method of singular value decomposition,[40] but it is much simpler to follow the lead of Sankey[17] and use shorter ranged basis functions instead. Here these were constructed by solving the (pseudo)atomic Schrödinger equation for the atom enclosed in a spherical box defined by adding the potential

$$V_{\rm box}(r) = \frac{a}{1 + e^{b(r_m - r)}} \quad (11)$$

to the atomic Hamiltonian. The values of a and b were fixed at 5 Hartrees and $10a_0^{-1}$, respectively, but a variety of different box radii, r_m, were tried. The Harris functional does not require any special relationship between the form of the basis functions and the form of the input density, and tests showed that it made little difference whether input densities were obtained by superposing boxed or unboxed atoms. Boxed densities were used in most of this work.

The largest box radius investigated was $r_m = 6a_0$. The presence of the box increased the pseudoatomic s and p valence eigenvalues from the free atom values of -0.42 Hartrees and -0.15 Hartrees to -0.39 Hartrees and -0.12 Hartrees, respectively, but neither the atomic eigenfunctions nor the calculated dimer interatomic spacing and vibrational frequency were much altered. Sankey's experience was that basis functions defined using such large box radii are quite adequate in the solid, but the band structure obtained here was nothing like the accurate Harris functional band structure. Since these calculations were properly converged as a function of the real space cutoff, the only conclusion is that the errors must be due to the two-center approximation, albeit magnified because the overlap matrix still has a few small eigenvalues. Equally large errors were seen in the total energy (calculated with the same 10 special k point set as was used in the plane wave calculations), which was found to decrease monotonically with the lattice parameter right down to 4.76 Å (the smallest value investigated).

As the box radius is decreased, the band structure changes smoothly until at $r_m = 4a_0$ it is as shown in Fig. 2. This is much closer to the correct band structure of Fig. 1, but the severe truncation of the basis functions spoils the energetics. The atomic s and p valence eigenvalues are now -0.26 Hartrees and 0.05 Hartrees and the dimer equilibrium spacing is $4.2a_0$. The calculated lattice parameter of the solid is 5.22 Å and the calculated bulk modulus is 1670 kbar. The pressure derivative of the bulk modulus (obtained from the self-consistent DF calculations) was 5.1, which implies $B \simeq 1410$ kbar when the lattice parameter is 5.22 Å, and so the error in the bulk modulus is at least consistent with the error in the lattice parameter. Adding the lowest d eigenfunctions of the boxed atom to the basis set improves the band structure a little, but the lattice parameter and bulk modulus do not change much. Adding additional excited s functions to the basis also makes little difference.

In summary, this section has demonstrated that the simple two-center approximation is not adequate in Ge. Atomic orbitals are not suitable as basis functions because they are too long ranged and close to being linearly

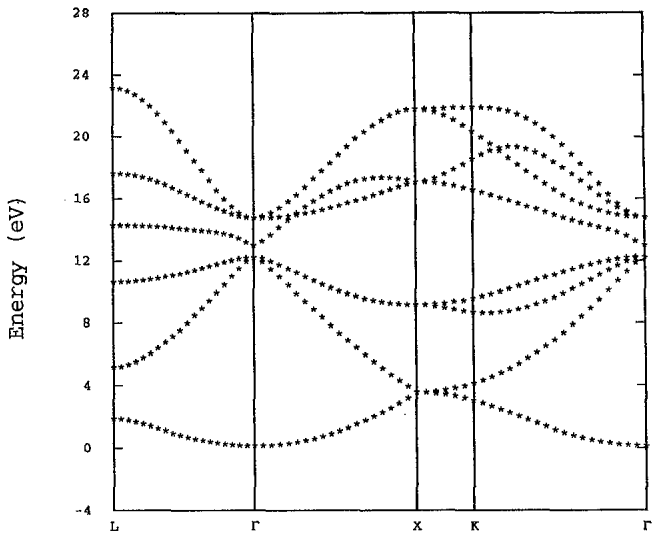

FIG. 2. Non-self-consistent Harris functional tight-binding Ge band structure ignoring three-center contributions to the Hamiltonian matrix. A basis set of boxed s and p (pseudo) atomic orbitals was used with the box size, r_m, equal to $4a_0$.

dependent, and these two factors combine to make the results sensitive to small H matrix elements between very distant neighbors. The problem can be cured by truncating the basis functions and Sankey[17] has shown that this yields a workable method as long as all the three-center terms are included. Unfortunately, the degree of truncation required to make the three-center terms unimportant is such that the basis set is no longer adequate.

V. TWO-CENTER TIGHT-BINDING AND CHEMICAL PSEUDOPOTENTIALS

Since it seems that the simple two-center approximation is inadequate in Ge, it is now necessary to consider the more sophisticated version based on CP theory.[25–28,20] This was originally developed as a solution to the problem of constructing a "perfect" basis set of localized functions exactly spanning a band of eigenstates in the solid. Wannier functions[41] are an example of such a basis set and are convenient because they are orthonormal, but they are not as localized as one would like and are difficult to construct. The main achievement of CP theory was to show that much more localized basis functions can be constructed if the orthonormality constraint is relaxed. These basis functions are the solutions of a self-consistent differential equation, the CP equation, which depends only on the local environment and can be solved without calculating all the eigenstates of the crystal first.

Consider the non-self-consistent Harris functional Hamiltonian of Eq. (8). The characteristic property defining a set of basis functions, $\{|\phi_1\rangle, \ldots, |\phi_M\rangle\}$, exactly spanning some particular group of M eigenfunctions, is that the set must be closed under the operation of $\hat{H}$,

$$\hat{H}|\phi_\alpha\rangle = D_{\beta\alpha}|\phi_\beta\rangle \quad , \quad (12)$$

where the summation convention applies for repeated suffixes. The point is that if Eq. (12) is satisfied, then any linear combination of the basis functions which diagonalizes the finite matrix D is an exact eigenfunction of $\hat{H}$. By taking matrix elements of Eq. (12) it can be seen that $D_{\alpha\beta} = S^{-1}_{\alpha\gamma} H_{\gamma\beta}$ and hence D is simply related to the familiar H and S matrices. The eigenvalues, usually calculated by solving $\det(H - \varepsilon S) = 0$, may also be found from $\det(D - \varepsilon I) = 0$. H and S are Hermitian matrices, of course, but D need not be Hermitian unless H and S commute.

At first sight, Eq. (12) does not appear very helpful. Everything becomes much clearer, however, if it is rewritten in the form

$$\hat{H}|\phi_\alpha\rangle - \sum_\beta G_{\beta\alpha}|\phi_\beta\rangle = \varepsilon^{(\alpha)}|\phi_\alpha\rangle \quad , \tag{13}$$

where $\varepsilon^{(\alpha)}$ is a constant still to be specified and $G_{\beta\alpha} \equiv D_{\beta\alpha} - \varepsilon^{(\alpha)}\delta_{\beta\alpha}$. The second term on the left hand side of Eq. (13) can be regarded as the result of the action of an operator on $|\phi_\alpha\rangle$. This operator has the property that no matter which state it acts on the outcome is entirely within the space of basis functions, and so it can be written in the form $\hat{P}\hat{G}$, where $\hat{P}$ is the projector on to the basis. Equation (13) then becomes

$$\left(\hat{H} - \hat{P}\hat{G}\right)|\phi_\alpha\rangle = \varepsilon^{(\alpha)}|\phi_\alpha\rangle \quad , \tag{14}$$

and the significance of the name "chemical pseudopotentials" becomes clear. Equation (14) looks very like the general defining equation for ordinary pseudopotentials given by Austin, Heine, and Sham,[42] except that here the projection operator involves the orbitals to be calculated instead of the core states.

The operator $\hat{G}$ is arbitrary and any choice for which Eq. (14) is soluble will generate a set of basis functions exactly spanning a group of eigenfunctions. The trick, of course, is to choose $\hat{G}$ to make the basis functions as localized as possible. Several suggestions have been put forward but Anderson's[27] is probably the simplest. He chooses

$$\hat{G} = \sum_\alpha \sum_{\beta(\neq\alpha)} |\phi_\beta\rangle\langle\phi_\beta|\hat{V}_\beta\hat{P}_\alpha \quad , \tag{15}$$

where $\hat{P}_\alpha$ is the projection operator on to the orbital $|\phi_\alpha\rangle$. Equation (14) then becomes

$$\left[\hat{T} + \hat{V}_\alpha + \hat{W} + \sum_{\beta(\neq\alpha)} \left(1 - |\phi_\beta\rangle\langle\phi_\beta|\right)\hat{V}_\beta\right]|\phi_\alpha\rangle = \varepsilon^{(\alpha)}|\phi_\alpha\rangle \quad , \tag{16}$$

which is known as the CP equation.

Near $\mathbf{R}_\alpha$ the pseudo-Hamiltonian is almost the same as the full Hamiltonian but on surrounding atoms the potentials have been reduced by the removal of the projections along $|\phi_\beta\rangle$ and the remaining weak "chemical pseudopotentials" do not have much effect. The solutions are therefore localized and look more or less like atomic orbitals. In Ge, where s and p (and to some extent d) orbitals are all important in the same energy range, subtracting the projection along just a single orbital on each site is not sufficient, but it is easy enough to generalize the theory to include several orbitals on each atom when required.

Once the basis functions have been found, the elements of the D matrix are given by

$$D_{\alpha\alpha} = \varepsilon^{(\alpha)} = \langle\phi_\alpha|\left[\hat{T} + \hat{V}_\alpha + \hat{W} + \sum_{\beta(\neq\alpha)} \left(1 - |\phi_\beta\rangle\langle\phi_\beta|\right)\hat{V}_\beta\right]|\phi_\alpha\rangle, \tag{17}$$

$$D_{\beta\alpha} = \langle\phi_\beta|\hat{V}_\beta|\phi_\alpha\rangle \quad . \tag{18}$$

Although CP theory provides an elegant solution to the problem of finding a perfect set of localized basis functions, its main use so far[20] has been to justify the use of atomic orbitals as basis functions. If the cancellation of the potentials on neighboring atoms is really as effective as Anderson's[43] calculations suggest, then the CP basis functions should look much like atomic orbitals ($\hat{W}$ is a weak potential, remember). Assuming this to be the case and substituting atomic orbitals into Eqs. (17) and (18), one sees that the D matrix has an almost perfect two-center form. CP theory suggests that making a two-center approximation to the D matrix may be sensible even though it has already been demonstrated that making a two-center approximation to H is not accurate.

But how good is this approximation? In conventional TB calculations, the basis functions are chosen at the beginning and their degree of similarity to the CP functions is hard to establish. All one can do is trust in the variational principle and there is a good deal of evidence[20,17,18] that this acts effectively enough in many cases. CP theory suggests that ignoring three-center terms in the D matrix may not spoil the variational accuracy as much as ignoring the three-center terms in the H matrix, but this is not obvious. An alternative approach to D matrix TB, discussed below, sheds some more light the matter. It leads to a slightly different two-center D matrix from the one obtained by substituting atomic orbitals into Eqs. (17) and (18), but the two approaches give similar results in practice.

First it is necessary to define the "variational" D matrix. Suppose that the fixed TB basis functions are denoted by $|\phi_{\alpha i}\rangle$. Then the D matrix defined by the equation

$$\hat{P}\hat{H}|\phi_{\alpha i}\rangle = \sum_{\beta j} D_{\beta j,\alpha i}|\phi_{\beta j}\rangle \tag{19}$$

(where $\hat{P}$ is the projector onto the fixed basis) satisfies $H = SD$. The D matrix secular equation, $\det(D - \varepsilon I) = 0$, is therefore equivalent to the familiar variational secular equation, $\det(H - \varepsilon S) = 0$, and the D matrix has real and variationally accurate eigenvalues. Equation (19) is clearly the best definition of the D matrix for fixed basis set calculations.

In order to see how to make a two-center approximation to this D matrix, define the atomic and dimer Hamiltonians, $H_\alpha = \hat{T} + \hat{V}_\alpha$ and $H_{\alpha\beta} = \hat{T} + \hat{V}_\alpha + \hat{V}_\beta + \hat{W}_{\alpha\beta}$ (where $W_{\alpha\beta}(\mathbf{r}) = \mu_{\rm xc}([n_\alpha + n_\beta], \mathbf{r}) - \mu_{\rm xc}([n_\alpha], \mathbf{r}) - \mu_{\rm xc}([n_\beta], \mathbf{r})$ is the dimer contribution to the nonspherical part of the exchange-correlation potential). Now write the left hand side of Eq. (19) in the form

$$\hat{P}\hat{H}|\phi_{\alpha i}\rangle = \hat{P}_\alpha \hat{H}_\alpha |\phi_{\alpha i}\rangle + \sum_{\beta(\neq\alpha)} (\hat{P}_{\alpha\beta}\hat{H}_{\alpha\beta} - \hat{P}_\alpha \hat{H}_\alpha)|\phi_{\alpha i}\rangle + \hat{\Delta}|\phi_{\alpha i}\rangle, \quad (20)$$

$$= \sum_{\beta j} \tilde{D}_{\beta j,\alpha i}|\phi_{\beta j}\rangle + \hat{\Delta}|\phi_{\alpha i}\rangle, \quad (21)$$

where $\hat{P}_\alpha$ is the projector onto the space spanned by all basis functions on atom α, and $\hat{P}_{\alpha\beta}$ is the projector onto the space spanned by all basis functions on atoms α and β. The first term on the right hand side of Eq. (20) is the one-center contribution, the second term is the sum of the two-center contributions, and $\hat{\Delta}$ contains everything involving three or more centers. The approximate two-center D matrix, $\tilde{D}$, therefore contains only one- and two-center matrix elements and makes only three-center errors.

Equations (20) and (21) define a two-center D matrix and show how all the matrix elements are to be evaluated from monomer and dimer calculations. The hope is that this should be better than making a two-center approximation to the H matrix, but how much better? One way to check is to examine the matrix elements of $\hat{\Delta}$. General expressions for these can be written down, but it is more instructive to consider the special case when the basis functions are atomic orbitals and to approximate by ignoring $\hat{W}$. All intra-atomic matrix elements of $\hat{\Delta}$ are then zero, and the inter-atomic matrix elements take the form

$$\hat{\Delta}_{\beta j,\alpha i} = \langle\phi_{\beta j}| \sum_{\gamma\neq\alpha,\beta} (1 - \hat{P}_{\alpha\gamma})\hat{V}_\gamma|\phi_{\alpha i}\rangle \quad . \quad (22)$$

These are all three-center terms as expected, with the potential on site γ weakened by the action of the projection operator $(1 - \hat{P}_{\alpha\gamma})$. The more basis functions on sites α and γ, the more effective this projection will be, and the smaller the three-center error will be in any given matrix element. Ignoring three-center terms in the D matrix is therefore equivalent to ignoring three-center matrix elements of weakened chemical pseudopotentials, and should indeed be better than making a two-center approximation to the H matrix (which ignores three-center matrix elements of the full atomic potentials).

VI. THE TWO-CENTER APPROXIMATION FOR THE D MATRIX

The arguments for two-center D matrix TB seem fairly convincing but how well does it do in practice? The large body of work by Bullett[20] shows that in many materials (transition metals in particular) it works very well, but unfortunately this does not seem to be the case for Ge. The following calculations employ the same basis sets and k points as were used in testing the simple two-center approximation in Sec. IV, and so all differences between the two sets of results can be attributed to the different forms of the two-center approximation.

Calculations using the $4s$ and $4p$ atomic orbital basis were not as sensitive to distant neighbor interactions as the H and S matrix calculations and the results were well converged with a real space cutoff of 7.4 Å. However, the calculated band structure was still nothing like the accurate Harris functional band structure. Truncating the basis functions using $r_m = 4a_0$ gives the band structure shown in Fig. 3, which is *much* worse than the corresponding HS matrix band structure. The calculated lattice parameter was 5.44 Å and the bulk modulus was 1440 kbar, to be compared to the accurate Harris functional values of 5.50 Å and 820 kbar, respectively. The band structure obtained when the five $4d$ orbitals were added to the $4s$ and $4p$ orbital basis set with $r_m = 4a_0$ was worse, if anything, than the band structure shown in Fig. 3. The D matrix band structures look so wrong that it is reasonable to ask whether there is an error in the calculations. This is possible, of course, but the programs were well tested and Bullett[44] has obtained similar results using an unrelated set of programs.

So where does the problem lie? The arguments in favor of two-center D matrix TB rely on the assumption that the three-center matrix elements of the CP are smaller than those of the bare potential. This seems sensible, but consider the simple case when there is only one basis function per atom and compare the magnitude of $\langle\phi_\beta|\hat{V}_\gamma|\phi_\alpha\rangle$ with the magnitude of the corresponding matrix element of the Anderson CP,

$$\langle\phi_\beta|(1 - |\phi_\gamma\rangle\langle\phi_\gamma|)\hat{V}_\gamma|\phi_\alpha\rangle = \langle\phi_\beta|\hat{V}_\gamma|\phi_\alpha\rangle - \langle\phi_\beta|\phi_\gamma\rangle\langle\phi_\gamma|\hat{V}_\gamma|\phi_\alpha\rangle \quad . \quad (23)$$

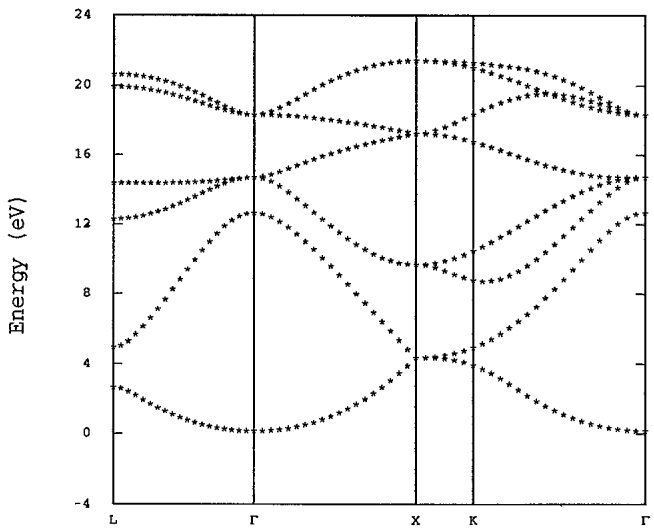

FIG. 3. Non-self-consistent Harris functional tight-binding Ge band structure ignoring three-center contributions to the D matrix. A basis set of boxed s and p (pseudo) atomic orbitals was used with the box size, r_m, equal to $4a_0$.

Since the atomic orbitals are longer ranged than the atomic potentials, the second term on the right hand side decays to zero more slowly than the first as the distance between atoms β and γ increases. For large values of $|\mathbf{R}_\beta - \mathbf{R}_\gamma|$, therefore, the second term may be much larger than the first. The three-center matrix element of the CP is then "overscreened" and may end up larger than the matrix element of the original potential (although of opposite sign), so ruining the accuracy of the two-center D matrix approximation. Neither the simple two-center H matrix approximation nor the more sophisticated two-center D matrix approximation works well for Ge, and neither explains the usefulness of two-center orthogonal semiempirical TB in tetrahedral semiconductors.

VII. CONCLUSIONS

The recent successes of the Harris functional used with very simple input densities constructed by superposing spherical atoms have led to the development of a number of parameter-free TB methods. These have been very successful, but would be much more appealing if a way could be found to avoid the evaluation of the many complicated three-center integrals. If this could be accomplished, then the parameter-free TB schemes would become almost as simple as semiempirical TB and many interesting applications would open up. The obvious approach is just to ignore all the three-center contributions to the H matrix, but this turns out to be a poor approximation for Ge. The three-center terms are not small enough to be neglected and the errors are magnified because of the near linear dependence of the basis set of atomiclike orbitals. The errors can be reduced by reducing the range of the basis functions, but the three-center terms remain important even when the orbitals are so short ranged that they no longer make a good basis set and so not much can be gained this way. The long accepted argument that CP theory is the answer, and that making a two-center approximation to the D matrix is accurate even though the two-center H matrix approximation is not, has been shown to be wrong in Ge. The three-center contributions to the D matrix are at least as important as the three-center contributions to the H matrix and cannot be neglected.

It seems, therefore, that the use of two-center TB for tetrahedral semiconductors (and presumably for all solids except those clearly in the conventional narrow band TB limit) is still on shaky foundations and that the promise of the D matrix approach is not fulfilled. Bullett's work has shown that this method works well when the bands are narrow, but the simple two-center H matrix approximation should also be good in such cases. When both methods are inaccurate, as in Ge, it seems that the D matrix approach is the poorer of the two and its possible applications are therefore very limited.

The question about why Goodwin's[24] simple two-center orthogonal TB model for Si worked so well remains. Is it just a sensible physically motivated fitting scheme, or do his parameters have precise physical meanings and values which could have been calculated using DFT? This work has shown that the conventional answers to these questions are inadequate and so some new ideas are required.

[1] I. Stich, M. C. Payne, R. D. King-Smith, J. S. Lin, and L. J. Clarke, Phys. Rev. Lett. **68**, 1351 (1992).
[2] K. D. Brommer, M. Needels, B. E. Larson, and J. D. Joannopoulos, Phys. Rev. Lett. **68**, 1355 (1992).
[3] J. K. Norskov and N. D. Lang, Phys. Rev. B **21**, 2131 (1980).
[4] M. J. Stott and E. Zaremba, Phys. Rev. B **22**, 1564 (1980).
[5] K. W. Jacobsen, J. K. Norskov, and M. J. Puska, Phys. Rev. B **35**, 7423 (1987).
[6] M. W. Finnis and J. Sinclair, Philos. Mag. A **50**, 45 (1984).
[7] M. S. Daw and M. I. Baskes, Phys. Rev. Lett. **50**, 1285 (1983); Phys. Rev. B **29**, 6443 (1984).
[8] A. E. Carlsson, P. A. Fedders, and C. W. Myles, Phys. Rev. B **41**, 1247 (1990).
[9] J. A. Moriarty, Phys. Rev. B **38**, 3199 (1988); J. A. Moriarty and R. Phillips, Phys. Rev. Lett. **66**, 3036 (1991).
[10] D. G. Pettifor, Phys. Rev. Lett. **63**, 2480 (1989).
[11] J. Harris, Phys. Rev. B **31**, 1770 (1985).
[12] W. M. C. Foulkes and R. Haydock, Phys. Rev. B **39**, 12 520 (1989).
[13] A. J. Read and R. J. Needs, J. Phys. Condens. Matter **1**, 7565 (1989).
[14] H. M. Polatoglou and M. Methfessel, Phys. Rev. B **37**, 10 403 (1988).
[15] M. W. Finnis, J. Phys. Condens. Matter **2**, 331 (1990).
[16] I. J. Robertson, M. C. Payne, and V. Heine, J. Phys. Condens. Matter **3**, 8351 (1991).
[17] O. F. Sankey and D. J. Niklewski, Phys. Rev. B **40**, 3979 (1989).
[18] A. Gibson, R. Haydock, and J. P. LaFemina, Phys. Rev. B **47**, 9229 (1993).
[19] J. Harris and D. Hohl, J. Phys. Condens. Matter **2**, 5161 (1990).
[20] For general reviews of semiempirical tight-binding and the chemical pseudopotential method, see the articles by D. Bullett, R. Haydock, V. Heine, and M. J. Kelly, in *Solid State Physics: Advances in Research and Applications*, edited by H. Ehrenreich and D. Turnbull (Academic, New York, 1980), Vol. 35.
[21] W. A. Harrison, *Electronic Structure and the Properties of Solids* (Freeman, San Francisco, 1980).
[22] D. J. Chadi, Phys. Rev. Lett. **43**, 43 (1979).
[23] A. T. Paxton and A. P. Sutton, Act. Met. **37**, 1693 (1989).
[24] L. Goodwin, A. J. Skinner, and D. G. Pettifor, Europhys. Lett. **9**, 701 (1989).
[25] W. H. Adams, J. Chem. Phys. **34**, 89 (1961).
[26] T. L. Gilbert, in *Molecular Orbitals in Chemistry, Physics and Biology*, edited by B. Pullman and P. O. Löwdin (Academic, New York, 1964).
[27] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[28] J. D. Weeks, P. W. Anderson, and A. G. H. Davidson, J. Chem. Phys. **58**, 1388 (1973).

[29] J. C. Slater and G. F. Koster, Phys. Rev. **94**, 1498 (1954).
[30] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[31] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[32] For a general review of density functional theory, see the article by C. J. Callaway and N. H. March, in *Solid State Physics: Advances in Research and Applications*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1984), Vol. 38.
[33] A. J. Skinner and D. G. Pettifor, J. Phys. Condens. Matter **3**, 2029 (1991).
[34] T. Starkloff and J. D. Joannopoulos, Phys. Rev. B **16**, 5212 (1977).
[35] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[36] J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[37] A. Keane, Austr. J. Phys. **7**, 323 (1954).
[38] M. T. Yin and M. L. Cohen, Phys. Rev. B **26**, 5668 (1982).
[39] J. Harris and R. O. Jones, Phys. Rev. A **18**, 2159 (1978).
[40] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1989).
[41] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt, Rinehart and Winston, New York, 1976).
[42] B. J. Austin, V. Heine, and L. J. Sham, Phys. Rev. **127**, 276 (1962).
[43] P. W. Anderson, Phys. Rev. **181**, 25 (1969).
[44] D. Bullett (private communication).