Unconstrained and constrained minimization, linear scaling, and the Grassmann manifold: theory and applications

David Raczkowski and C.Y. Fong
Department of Physics, University of California, Davis, CA 95616-8677

Peter A. Schultz and R.A. Lippert
Sandia National Laboratories, Albuquerque, NM 87185-1413

E.B. Stechel
Ford Motor Co. srl md 3028, Dearborn, MI 48211-2053

Abstract

An unconstrained minimization algorithm for electronic structure calculations using density functional for systems with a gap is developed to solve for nonorthogonal Wannier-like orbitals in the spirit of [E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B 49, 10 008 (1994)]. The search for the occupied sub-space is a Grassmann conjugate gradient algorithm generalized from the algorithm of [A. Edelman, T. A. Arias, and S. T. Smith, SIAM J. on Matrix Anal. Appl. 20, 303 (1998)]. The gradient takes into account the nonorthogonality of a local atom-centered basis, gaussian in our implementation. With a localization constraint on the Wannier-like orbitals, well-constructed sparse matrix multiplies lead to O(N) scaling of the computationally intensive parts of the algorithm. Using silicon carbide as a test system, the accuracy, convergence, and implementation of this algorithm as a quantitative alternative to diagonalization are investigated. Results up to 1458 atoms on a single processor are presented.

71.10.+x 71.20.Ad

I. Introduction

In the past decade, numerous minimization techniques have appeared in the condensed matter literature for solving for the ground state in electronic structure calculations.[1-19] A common bottleneck in Hartree-Fock (HF), density functional theory (DFT), and tight-binding (TB) methods has been the $O[N^3]$ scaling in the computational effort required to generate the ground state solution, where N is proportional to the number of particles in the system. For systems with an energy gap, e.g. insulators, semiconductors, or molecules, minimization techniques offer methods to calculate the charge densities and total energies with computational effort that scales linearly, O(N), with the size of the system.

Minimization techniques achieve linear scaling by taking advantage of well-known chemical intuition recently summarized in the "near-sightedness" principle.[19]

DISCLAIMER

Two sufficiently separated regions of a molecule, bulk system, surface, etc. should not interact strongly. Implicit in this principle is a localization condition: what is a range beyond which interactions can be neglected? This is not a deterministic question, but is a balance between accuracy and computational efficiency. Smaller localization regions (see Sec. V) imply longer range interactions are truncated, leading to reduced accuracy, but lead to computational savings as fewer matrix elements need be computed. Thus, the O(N) minimization techniques have mostly been implemented in TB methods, where the interaction are shorter ranged (cf. HF and DFT). The resulting matrices are sparser, leading to an earlier crossover point to the regime where linear scaling minimization techniques are more efficient than diagonalization to obtain the electronic ground state.

The basis sets of DFT and HF are larger and longer range than in TB, leading to Hamiltonian (or Fock) matrices that are larger and much less sparse. The longer range, and more computationally demanding, interactions of HF and DFT postpone the crossover point of O(N) minimization compared to diagonalization, limiting the benefits of linear scaling algorithms for finding the ground state solution. The crossover point may be achieved at smaller system sizes through smaller localization regions but only at a cost of accuracy. Hence, whether it is advantageous to substitute a linear scaling minimization for explicit diagonalization depends not only on system size, but also on the degree of accuracy desired in the solution. In this paper, we will examine the question of how desired accuracy affects the crossover point at which the linear scaling algorithms become more efficient than diagonalization in DFT calculations.

We present the implementation of a Grassman conjugate gradient (GCG) minimization method into a Gaussian-based DFT code. We use non-orthogonal orbitals to span the occupied space, and discuss the technical issues related to the practical implementation of the minimization algorithm with localization. Using this implementation, we investigate the natural length scales of localization regions. The tradeoff between accuracy and computational efficiency and scaling behavior is discussed in detail. The method is demonstrated in full basis calculations of SiC bulk systems containing up to as many as 1,458 atoms and 18,954 basis functions.

Sec. II gives some mathematical background for the reader not familiar with the minimization algorithms. Sec. III introduces the common geometric framework of the minimization techniques: the Grassmann manifold.[20,21] The aim of this section is to present this recent mathematical insight in a concise, accessible manner and to add new perspectives that pertain to linear scaling. Sec. IV presents our minimization algorithm, which has application to any problem that requires the sum of any number of the lowest eigenvalues of a standard or generalized eigenvalue problem and the sub-space spanned by the corresponding eigenvectors. Effort is especially made to ease conceptual understanding while maintaining mathematical rigor. The description of vector spaces is general and deals with possible nonorthogonality at every level. Sec. V relates how localization of the physical system at different levels translates into linear scaling of the minimization algorithm. We also present information specific to the linear scaling implementation of the minimization algorithm. Sec. VI gives results and numerical analysis for our test system, silicon carbide. Our intent of this section is to evaluate the ability of an implementation of the algorithm to obtain quantitatively usable relative energies. We investigate this ability by mapping out the accuracy of different localization regions compared to diagonalization.

II. Mathematical Background

The ground-state total energy, E_T, and charge density, $n(\mathbf{r})$, of a molecule or condensed matter are fundamental quantities of a system.[22] Within an independent electron picture, DFT is a common method to calculate these quantities. There are two main variational formulations for these calculations – density matrix[12-19] and orbital[1-11] approaches. Both give the same total energy and charge density but obtain them in slightly different ways. With regards to notation, bold face will be used for vectors and matrices.

The orbital formulation for systems with a gap traditionally solves a generalized matrix eigenvalue problem

$$\mathbf{H\Psi=S\Psi E}. \tag{1}$$

For a given representation (basis set of M functions), $\mathbf{H}$ is the M x M Hamiltonian matrix. $\mathbf{S}$ is the overlap matrix, and equals $\mathbf{I}$, the unit matrix, in an orthonormal basis. Ψ is the M x N matrix comprised of the expansion coefficients of the M basis functions for the N lowest normalized eigenvectors of $\mathbf{H}$. Ψ diagonalizes $\mathbf{H}$, creating the diagonal N x N matrix $\mathbf{E}$, and defines the *occupied* vector space. In general, the eigenvectors are delocalized, making Ψ a dense matrix.

When calculating E_T and $n(\mathbf{r})$ of systems with a gap, all eigenvectors are weighted equally. Thus, only the collective properties of the eigenvectors are necessary to obtain E_T and $n(\mathbf{r})$. We define $\Psi(\mathbf{r})$ to be the projection of the *occupied* space onto real space; therefore, if we approximate real space by a mesh of 100 points, $\Psi(\mathbf{r})$ would be a matrix of 100 x N. One can obtain E_T from the band energy, E_B, and $n(\mathbf{r})$ using

$$E_B = \mathrm{Tr}[\mathbf{E}] \tag{2}$$

and

$$n(\mathbf{r}) = \text{diagonal matrix elements of } [\Psi(\mathbf{r})\,\Psi^{\dagger}(\mathbf{r}')], \text{ thus } \mathbf{r}=\mathbf{r}'. \tag{3}$$

This formulation alleviates the burden of orthogonalization of the wavefunctions as follows. For transformations $\Psi\mathbf{A} = \Phi$, with $\mathbf{A}$ being any N x N non-singular matrix, Φ spans the same space as Ψ. E_B and $n(\mathbf{r})$ are invariant as long as the equations are generalized to handle nonorthogonality and lack of explicit normalization. The essential information, E_B and $n(\mathbf{r})$, traditionally obtained by diagonalization, are now obtained in a different manner that allows for O(N) scaling for the solution of the ground state.

For the N lowest eigenfunctions, the matrix eigenvalue problem is formally equivalent to minimizing the trace of a matrix Rayleigh quotient equation[23]

$$E_B(\Phi) = \mathrm{Tr}[(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\,\Phi^{\dagger}\mathbf{H}\Phi\,] \tag{4}$$

with respect to Φ, under the constraints that $\Phi^{\dagger}\mathbf{S}\Phi = \mathbf{I}$, and $\Phi^{\dagger}\mathbf{H}\Phi = \mathbf{E}$ is a diagonal matrix.[7,8] As the total electronic energy and charge density of the entire system are

desired and not of individual states, the removal of these constraints leaves E_B unaltered as long as $\Phi^\dagger S\Phi$ is not a singular matrix. Φ is nonorthogonal, and most importantly for our purposes, made local (see Sec. V). The charge density, n(r), is also unaltered when the following formula is used:

$$n(\mathbf{r}) = \text{diagonal matrix elements of } [\Phi(\mathbf{r})(\Phi^\dagger S\Phi)^{-1}\Phi^\dagger(\mathbf{r}')].^{24} \quad (5)$$

The density matrix formulation also obtains information only about the entire system and not individual states. The band energy, Eq. (6), is minimized with respect to the density matrix **P**, a hermitian M x M matrix:

$$E_B(\mathbf{P}) = \mathrm{Tr}[\mathbf{PH}]. \quad (6)$$

As $\mathbf{P} = \Phi(\Phi^\dagger S\Phi)^{-1}\Phi^\dagger$, and from the relation Tr[**AB**] = Tr[**BA**], we can see that Eq. (4) and Eq. (6) are equivalent. The matrix **P** is obtained directly without having to calculate Φ. However the idempotent constraint **PSP=P**, automatic when $(\Phi^\dagger S\Phi)^{-1}$ is calculated, must be achieved for the ground state solution. This is usually done by utilizing (see Sec. III) the Mcweeny purification:[25]

$$3\mathbf{PSP} - 2\mathbf{PSPSP} \rightarrow \mathbf{P}. \quad (7)$$

III. Geometry of the Vector Space: the Grassmann Manifold

The Grassmann manifold[20, 21] of rank N is the set of all subspaces of rank N in some ambient (*primitive*) M dimensional space. There are two common representations used in physics:

Density Matrix: **P** - a hermitian M x M matrix with the idempotent constraint **PSP=P**

Orbital:

Orthogonal Ψ - a M x N matrix with the orthogonality constraint of $\Psi^\dagger S\,\Psi = \mathbf{I}$

Non-orthogonal Φ - a M x N matrix with the constraint that $(\Phi^\dagger S\Phi)^{-1}$ is non-singular.

In the ground-state density matrix formulation, the *primitive* space is the M(M+1)/2-dimensional space of all hermitian M x M matrices. Each matrix is a point in this space and defines the occupation magnitude for 2N electrons, the *occupied* space. To account for any possible hermitian M x M matrix, the occupation magnitude could be any non-negative value, most of which give unphysical solutions. A ground-state density matrix must satisfy idempotency. The Grassmann manifold, idempotent surface[15], corresponds to the points that satisfy this condition.

The description for the orbital formulation is more involved. For notational and conceptual reasons, we rearrange Eq. (4) by defining

$$\overline{\Phi}^{\dagger} = (\Phi^{\dagger} S \Phi)^{-1} \Phi^{\dagger} \tag{8}$$

which gives

$$E_B(\overline{\Phi}^{\dagger}, \Phi) = \mathrm{Tr}[\overline{\Phi}^{\dagger} H \Phi] \tag{9}$$

as our functional to be minimized. Here we choose to have $\overline{\Phi}^{\dagger}$ as our variable instead of $\overline{\Phi}$ as it is always the transpose that appears in our equations. Eq. (9) is in a dual basis[26-28] form where $\overline{\Phi}^{\dagger}$ is the covariant matrix, one-form[29] or linear-form[30], of the matrix Φ. Biorthogonality, of which orthogonality is a special case, is automatically satisfied if the inverse is calculated in Eq. (8)

$$\overline{\Phi}^{\dagger} S \Phi = (\Phi^{\dagger} S \Phi)^{-1} \Phi^{\dagger} S \Phi = \mathbf{I}. \tag{10}$$

The points in space are still defined by our occupied space, but the occupied space is now defined by $(\overline{\Phi}^{\dagger}, \Phi)$ and is related to **P** through

$$\mathbf{P} = \Phi \overline{\Phi}^{\dagger}. \tag{11}$$

If $\overline{\Phi}^{\dagger}$ and Φ are biorthogonal, then **P** is idempotent and Tr[PS]=2N; therefore, the point (**P** or $(\overline{\Phi}^{\dagger}, \Phi)$) resides on the Grassmann manifold. Given a nonsingular M x M matrix A and $\overline{\Phi}_A^{\dagger}$, the biorthogonal complement of Φ **A**, $(\overline{\Phi}^{\dagger}, \Phi)$ and $(\overline{\Phi}_A^{\dagger}, \Phi \mathbf{A})$ through Eq. (11) create the same **P** and thus define the same point on the Grassmann manifold.

An equivalent perspective defines each point by Φ and a covariant metric of the occupied subspace creating $\overline{\Phi}^{\dagger}$ from Φ. If the correct metric is used, the point lies on the Grassmann manifold. It is in this space, Φ and a metric, that we minimize $E_B(\Phi)$, with respect to Φ, under the constraint that the minimum lies on the Grassmann manifold. The reason for introducing a constraint manifold into a previously unconstrained problem is that, in an asymptotically linear scaling algorithm, the exact covariant metric of the occupied subspace, $(\Phi^{\dagger} S \Phi)^{-1}$, is not calculated. Algorithms must address the possible departure and return to the manifold.

Satisfying Constraints

The task of bringing the density matrix to reside on the Grassman manifold, idempotency surface, after each update of **P** has been described as translations, a Mcweeny path,[15] due to repeated iterations of Eq. (7). Using sparse multiplications, this is an O(N) process. If Eq. (7) is used only to replace **P** inside of Eq. (6), creating a new functional, the path will ideally move in close proximity to the Grassman manifold and the minimum will be found on the manifold. If the path wanders too far away from the manifold, some eigenvalues of **P** may diverge to $-\infty$ or $+\infty$.[12]

Satisfying biorthogonality, adhering to the Grassmann manifold, for system sizes (see Sec. VI) where $\Phi^{\dagger} S \Phi$ is not appreciably sparse (see Sec. V) is straightforward as it is most efficient to calculate the dense $(\Phi^{\dagger} S \Phi)^{-1}$. Eventually the asymptotic $O(N^3)$ scaling of the dense inverse will dominate. The methods for satisfying the constraint with linear scaling effort are conceptually very similar to the density matrix methods. Some

methods[2-5] create a new functional that also allows the path to wander in close proximity to the surface. The final orbitals are orthogonal. Other proposals use transformation iterations[21] on Φ or a direct calculation[1] of an approximate $(\Phi^{\dagger}S\Phi)^{-1}$ for remaining on the Grassmann manifold after each update of Φ.

It is not plainly evident which method is superior. A nonorthogonal orbital method might be preferable over orthogonal methods given the observation that nonorthogonal orbitals are typically more localized.[31,32] The convergence rate for the minimization is slower for the algorithms that are allowed to wander off of the Grassman manifold.[33] This effect was only studied for dense linear algebra, but the effects should carry over to sparse linear algebra.

IV. Grassmann Conjugate Gradient Algorithm

Within the orbital formulation, a Grassman conjugate gradient algorithm[20, 21] is used to minimize Eq. (4). Initially, we introduce the algorithm without the complexity of localization, i.e. no sparse matrices. The algorithm is exactly the same as described in Ref. [21] with the addition of the parallel translation of the gradient.

A. Properties of a Gradient

We need to define a gradient, $\mathbf{G}=\nabla_{\Phi} E_B(\Phi)$, to be used in producing new conjugate search directions to minimize $E_B(\Phi)$. In [21], it was noted that the gradient of a functional should not be confused with its differential. The gradient must be in the direction of greatest change of the functional. Even an infinitesimal movement in the direction of the gradient must cause E_B to change in value; therefore, the gradient must have no component along Φ because such a direction would not change the value of E_B. This is a property of the functional and does not depend upon the representation. For this to be true, the gradient must lie in the tangent plane[20] of Φ, a constraint given by

$$\Phi^{\dagger}S\mathbf{G} = 0 \text{ and equivalently } \overline{\Phi}^{\dagger}S\mathbf{G} = 0. \tag{14}$$

A gradient requires a well-defined inner product. For an inner product $\langle \bullet \rangle$ and an infinitesimal step s in the direction of a displacement $\delta\mathbf{V}$ in the tangent plane, the inner product of the gradient with $\delta\mathbf{V}$ must equal the directional derivative in $\delta\mathbf{V}$, given by

$$\langle \mathbf{G} \bullet \delta\mathbf{V} \rangle = \delta E_B = d/ds\, E_B(\Phi + s\delta\mathbf{V})|_{s=0}. \tag{15}$$

The only form that keeps the inner product invariant between arbitrarily complete representations, for vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ defined for Φ at the origin, is

$$\langle \mathbf{X}_1 \bullet \mathbf{X}_2 \rangle = \text{Real}\{\text{Tr}[(\Phi^{\dagger}S\Phi)^{-1}\mathbf{X}_1^{\dagger}S\mathbf{X}_2]\}. \tag{16}$$

This gives $\langle \Phi \bullet \Phi \rangle = N$, conserving electron number.

B. Gradient, Search Direction, and Line Minimization

The differential of the functional[33]

$$dE_B/d\Phi = (I - S\Phi\,\overline{\Phi}^{\dagger})H\Phi\,(\Phi^{\dagger}S\Phi)^{-1} \tag{17}$$

does not satisfy Eq. (14). The gradient

$$G = (S^{-1} - \Phi\,\overline{\Phi}^{\dagger})H\Phi \tag{18}$$

does satisfy Eq. (14). The differential, Eq. (17) lies on a cone around the gradient. This can be graphically seen in Fig. 3 of White et al.[34] It may provide search directions that are suitable for minimization, but the convergence will be degraded, as the gradient is not used. For efficiency, the differential may be preferable if incorporating S^{-1} is too expensive. This may be the case for finite elements[35,36] or Mehrstellen[37] finite difference representations, where the dimension of S is typically larger than for an LCAO basis.

The new search direction, Z_{I+1}, was updated by the Polak-Ribière (PR) formula[20],

$$Z_{I+1} = G_{I+1} + [\,\langle\,(G_{I+1} - G_I) \bullet G_{I+1}\rangle\ /\ \langle G_I \bullet G_I\,\rangle\,]\,Z_I. \tag{19}$$

This formula gives slightly better convergence than the Fletcher-Reeves (FR) form. A step size, Λ, in the search direction, **Z**, is chosen to minimize the functional in that direction. A quadratic approximation is used around the current point Φ for the step size of Λ in the direction **Z**.

$$E_B(\Phi + \Lambda Z) = E_B(\Phi) + \Lambda\, Z \bullet dE_B/d\Phi + \tfrac{1}{2}\,\Lambda\, Z \bullet H \bullet \Lambda\, Z \tag{20}$$

The second term is just the normal dot product of **Z** with the differential. The last term here is the matrix element, H (**Z**,**Z**) of the Grassmann Hessian. The general formula of the matrix element of the Grassmann Hessian, given two tangent vectors V_1 and V_2, is

$$H(V_1, V_2) = \mathrm{Tr}[\,\overline{V}_1^{\dagger}HV_2 - \overline{V}_1^{\dagger}SV_2\ \overline{\Phi}^{\dagger}H\Phi\,].^{40} \tag{21}$$

The overall cost of the calculation of H (**Z**,**Z**) is equivalent to the calculation of the gradient because the expensive calculation of **HZ** and **SZ** are already necessary for the next functional evaluation.

To get Λ, set:

$$dE_B(\Phi + \Lambda Z)/d\Lambda = 0. \tag{22}$$

The step size is now easily calculated by using Eq. (20), obtaining

$$\Lambda = Z \bullet dE_B/d\Phi\ /\ H(Z,Z), \tag{23}$$

and Φ is updated according to

$$\Phi_{NEW} = \Phi - \Lambda Z . \quad (24)$$

C. Parallel Transport and Convergence

Since the search direction should stay in the plane tangent to Φ, the current Z needs to be orthogonal to Φ_{NEW} so that when Eq. (19) is used in the I+1 GCG iteration Z_{I+1} will be orthogonal to Φ_{NEW}. In this manner, Hessian information will be properly communicated from one iteration to the next. This is accomplished to all orders by a parallel transport of Z corresponding to the update in Φ. G_I should also be translated. In Eq. (19), Z_I^{NEW} and G_I^{NEW} replace Z_I and G_I respectively.

$$Z_I^{NEW} = Z_I + \Lambda \Phi \overline{Z}_I^{\dagger} S Z_I \quad (24)$$

$$G_I^{NEW} = G_I + \Lambda \Phi \overline{Z}_I^{\dagger} S G_I \quad (25)$$

The convergence of the algorithm can be tracked in two ways. The algorithm can be terminated when the change in $E_B(\Phi)$ becomes smaller than a prescribed threshold. Termination can also occur once the norm of the gradient, $\mathrm{Real}\{\mathrm{Tr}[(\Phi^{\dagger} S \Phi)^{-1} G^{\dagger} S G]\}$ dips below a certain threshold. We have chosen the latter. In self-consistent field (SCF) calculations, one must update the Coulomb and the exchange-correlation potential by the charge density of the new orbitals. We have chosen to do this only after the norm of the gradient becomes sufficiently small for a given potential. For dense matrices a value of 10^{-10} and a maximum iteration number of 15 was sufficient to obtain results within a few μRy of diagonalization.

This orbital minimization method has utility in any eigenvalue problem that extensively uses iterative solvers. In electronic structure, this comprises representations that use a much larger number of basis functions than occupied states, e.g. plane waves.[7, 10, 11] finite difference,[37-39] and finite elements.[40,41]

V. Sparse Implementation of the Minimization Algorithm

We use a Gaussian-based LCAO method implemented in the serial code SEQQUEST[42] as the framework to implement the Grassmann conjugate gradient (GCG) minimization algorithm. Pseudopotentials enable concentration only on the valence orbitals, and a split-valence double zeta with polarization (DZP) basis is used. This basis set is optimized for accuracy without explicit consideration of sparsity in H and S.

A. Sparse Storage

The implementation of linear scaling requires using a basis set whose elements are strictly localized in real space, e.g. finite elements[40,41] or a tight-binding basis[2, 6], or pseudo-localized in real space, e.g. gaussians made local by the use of cutoffs. Using a local basis, H and S become sparse as a matrix element is exactly zero or set to zero once a "sufficient" separation distance between basis functions is reached. The specifics of O(N) calculation of H has been discussed elsewhere[43-48] and will not be repeated here.

Sparse storage of M x M matrices **H**, **S**, and **P** was easily adapted from the linear scaling dense creation scheme.[48] The cutoff values that determine if elements are non-negligible give errors in the total energy less than 1 μRy/atom. The sparsity pattern, a list of the positions of non-zero elements, of **S** is used for the storage of surviving elements of **H** and **P**.

The M x N matrices, Φ, **Z**, and **G**, have strict distance cutoffs input by the user. Only gaussians within the localization sphere, measured from the center of an orbital (atom-centered or bond-centered) contribute to that orbital. The option is available to only select certain gaussians from an atom. The localization radii for the contribution of the s-shells, p-shells, and d-shells can differ, as each shell has a different spatial extent.

The sparsity pattern of the M x N matrices of the type $S\Phi$ can be determined two ways. The first method is by use of cutoffs as for Φ. In the second method, an element is kept if its value is above an input threshold value. In the results presented in Sec. VI, the sparsity pattern was calculated and held fixed for each SCF cycle. Since the initial estimate of Φ was not sufficient for the second method, the initial SCF cycle used the first method and subsequent cycles used the second method.

Φ becomes sparse first while still allowing **P**, used in the calculation of the total energy and charge density, to be quite extended. **H**, **S**, and **P** sparsify next. $\Phi^{\dagger}S\Phi$ and $\Phi^{\dagger}H\Phi$ become sparse last due to the interaction of the occupied orbitals mediated by **H** and **S**. Since sparsity occurs at different stages it is advantageous for the algorithm to exploit the sparsity at different stages. If the matrices are sparse, the matrix multiplications can be carried out in O(N) steps by calculating only the elements that are non-zero. A matrix needs to be significantly sparse, about 10% of the elements being non-zero, before sparse routines become faster than machine-dependent optimized dense routines. For the system sizes studied, the sparsity of $\Phi^{\dagger}S\Phi$ and $\Phi^{\dagger}H\Phi$ did not warrant sparse multiplications; therefore, they were kept dense. A similar observation was noted in Ref. [9].

B. Action of S^{-1}

With localization, the definition of the gradient, Eq. (18), changes. If the full S^{-1}, an $O(N^3)$ process, is calculated, then the result of Eq. (18) is truncated in order for **G** to lie in the same space as Φ. This straightforward but naïve approach does not give the best results.

Our solution accounts for what the action of S^{-1} needs to accomplish when the orbitals are localized. With no localization, S^{-1} accounts for the curvature of the entire *primitive* space in order to align the differential in the direction of greatest increase. With localization, these orbitals are restricted to lie in a certain subspace of the entire vector space of our basis set. An orbital is affected only by the geometry of the vector space in its localized region. Therefore, only part of the overlap matrix, **S**, is associated with an orbital. This part of **S** is inverted in order to obtain the localized gradient for a given orbital.

First we retrieve the square submatrix of **S** corresponding to the overlap of the basis functions which are allowed to contribute to an orbital. This matrix is inverted and the result is stored in memory. The action of S^{-1} is then just a matrix multiply of a column

of Φ, a single localized orbital, by its corresponding overlap inverse, the contravariant metric of the subspace in which the orbital resides. This method scales linearly.

C. Problems with Convergence

With our implementation of the GCG algorithm including the effects of localization, the norm of the gradient no longer converges to zero. We do not know if this effect is a natural consequence of localization or our lack of fully understanding the geometry of the new vector space of localized orbitals. As long as the convergence is consistent and to a small enough value, the algorithm is reliable.

The norm of the gradient can not be used as the stopping criteria. The convergence is measured by

$$\mathbf{G} \bullet dE_B/d\Phi \tag{26}$$

Without localization, this is exactly the norm of the gradient.

The step size sometimes gives an increase in the energy. This occurs more often with small localization values. A version of Brent's algorithm has been implemented to ensure a decrease in E_B. For small localization regions, sometimes a decrease in E_B cannot be found in the search direction so the GCG algorithm is restarted.

VI. Results for Silicon Carbide

We test the accuracy of the GCG algorithm to resolve relative energies between different systems with different localization regions. Silicon carbide, a wide gap semiconductor that can be operated at high temperature and high pressure, was chosen for its technological importance[14] and for its multiple crystal phases. The latter provides a stringent test on the accuracy of the total energy calculations. The relative energy differences between the 3C (cubic) and the 2H and 4H (hcp) phases[49] are used for this purpose.

As a check of the basis set, Table 1A shows a favorable comparison of our converged calculations using diagonalization within the local density approximation (LDA) to structural values from experiment[50] and recent relative energy calculations using plane waves[49] (LDA) and LMTO[51] (GGA). We relaxed the internal positions for the hcp phases. The lattice constants and c/a ratios of the hcp phases lie within 1% of experiment. We obtain the proper energetic ordering of the phases, on the order of 0.1 mRy/atom.

For the two-atom 3C phase, we use up to a 12x12x12 Monkorst-Pack[52] mesh. We include the Γ point, $\mathbf{k} = 0$, for the **k**-point sampling, in order to obtain the total energy for a Γ point sampling of the Brillouin zone for the large unit cells. As the unit cell increases, the **k**-points of the primitive cell fold into each other and all eventually fold into the Γ point. For example, a calculation using a 5x5x5 mesh including the Γ point of the two-atom primitive cell of 3C SiC gives the same energy/atom as a 250-atom Γ point calculation. For the 4-atom 2H and 8-atom 4H, we used up to a 9x14x14 and a 3x10x10 mesh respectively. Table 1B gives the energy/atom for the different phases at different

system sizes obtained from k-point sampling as mentioned above. We will use this data to compare the accuracy of different localization regions.

Now imposing varying localization, we investigated the accuracy, convergence, and scaling of computational effort with system size for Γ point calculations of 64 to 1458 atoms. The initial occupied orbitals were formed from sp^3 hybridizing the single zeta gaussians, formed by fitting an atomic pseudo-wavefunction. The determination of the required accuracy, i.e. the energy differences to be resolved, and the corresponding localization region is logically the first step in these calculations. Generally, if the localization region is increased, higher accuracy is obtained. The localization also affects directly the convergence and computational effort. The desired accuracy also determines acceptable convergence criteria and growth of $S\Phi$, which also affects computational effort.

We started with the 64-atom system. Energies were calculated with the relaxed geometry from the converged k-point calculations. $S\Phi$ had full growth. Table 2 gives the energies for the different phases for a given localization including the diagonalization result for reference and the corresponding energy difference with the 2H phase. The localization is given as the cutoff radius for the single zeta, double zeta, and polarization shells. The double zeta shell is a single diffuse gaussian for the s and p shells. A "b" corresponds to an orbital whose localization region is centered between two atoms and "a" is for an atom-centered orbital. For example, a {5,4,5;b} setting centered in the bond, with the radius of the single zeta and polarization functions being longer than the double zeta. Four atom-centered orbitals on carbon were used for the larger localization region as the code takes advantage of shared sparsity patterns among orbitals to use less memory and run faster. At a 7 bohr radius cutoff, the code was more accurate and efficient for atom centered orbitals.

In order to have an equitable comparison with diagonalization and varying localization, the energy values in Table 2 are given after 4 SCF iterations and a maximum of 15 GCG iterations with a stopping criterion at 10^{-10} for Eq. (26). After 4 SCF iterations, the diagonalization calculations were converged to within 0.1mRy/atom. In order to take advantage of the smallest localization region for a desired accuracy, we looked at the effect that different gaussian functions had on the energy. Near the center of the orbital, the single zeta functions have a larger impact on the accuracy of the total and relative energies than the more diffuse double zeta functions. The single zeta functions give more variational freedom where the orbital has the highest probability and the strongest interactions, i.e. near the center of the bond.

With the localization radius between 5-9 bohr, having the single zeta with a larger radius than the double zeta gave lower energies energy for the 3C and 2H structures, but higher for 4H. In this range, a larger radius for double zeta gave better accuracy for all energy differences. Above 9 bohr, the double zeta gave lower energies for all structures, but the accuracy was mixed. Accuracy of less than 0.2 mRy/atom, which is on the order of the energy differences for the converged results, was obtained when the radius was 9 bohr for each shell. The 9 bohr radius corresponds to the extent of the orbitals of silicon determined by the cutoffs already in Seqquest, which give an accuracy of 1 μRy/atom. A radius of 12 bohr gave an accuracy of less than 0.1 mRy/atom, which also was obtained with a 9 bohr radius but only after 10 SCF iterations. The SCF and GCG convergence with a larger radius for the double zeta did generally better.

As the measure of convergence for the GCG algorithm, we plot, in Fig. 1, the absolute value of Eq. (26) for the 5,4,5-b setting of the 64-atom system for the first SCF iteration. Fig. 1 shows three distinct stages for the convergence of the GCG algorithm: (1) starts with linear convergence typical of a conjugate gradient algorithm, (2) hits a region of flat convergence, and (3) then ends with linear convergence. This behavior is exhibited in all localization regions. The duration of each stage and the corresponding transition points vary with localization radius. The good convergence at the end is generally accompanied by Eq. (26) being negative at some steps. In Fig. 1, Eq. (26) becomes and stays negative at iteration number 39. This behavior occurs for subsequent cycles and is explained as follows.

Initially, the largest gradient values are for the gaussians near the center of the localized orbital, which are within the localization region. This corresponds to the good initial convergence that results from the orbitals being allowed to change in the direction (gaussians) of steepest descent. As the gradient gets small, the size of the gradient for the gaussians within the localization region becomes comparable to the size for gaussians outside of the localization region. The error in the gradient due to localization is now of the same order of magnitude as its length. Thus, the GCG convergence stalls, as orbitals are not allowed to move in directions that would give a significant decrease in the energy. Localization has effectively cut off the bottom of the bowl that makes the minimum. The orbitals wander around the edge of this cut, as the direction that would take Φ to the bottom is no longer allowed. At the end, the sharp drop is most likely due to the two vectors $\mathbf{G}$ and $dE_B/d\Phi$ becoming perpendicular. $\mathbf{Z}$ and $dE_B/d\Phi$ also become perpendicular at this stage.

As convergence with localization differs from the dense case and since accuracy depends on the convergence, convergence criteria require investigation. Because the stopping criterion of 10^{-10} for Eq. (26) is rarely met in a reasonable number of iterations, the maximum number of iterations is more crucial. Generally the maximum of 15 iterations achieves the same accuracy as larger maximums. In some cases better accuracy is obtained with more iterations, but it is probably more efficient to use a larger localization in order to obtain the same accuracy. There is a balance between the accuracy obtainable and the effort expended due to a large number of GCG iterations.

A suitable stopping criterion for SCF convergence also depends on the localization. A larger localization region will achieve SCF convergence comparable to diagonalization, and a smaller localization may never achieve the stopping criteria used for diagonalization calculations. One solution to this problem is to know the accuracy obtainable for a given localization. Since less accuracy is expected from smaller localization regions, the SCF convergence can be stopped for a larger energy difference than for diagonalization and larger localization regions. For example, one can consider the {5,4,5;b} converged once the energy change is less than 1 mRy/atom as higher accuracy is not expected. This was achieved in 4 SCF iterations. The {9,12,9;a} converged to within 0.1mRy/atom after 4 SCF iterations. If the SCF procedure takes too long for the desired convergence, then one needs larger localization.

We now turn our attention to the growth parameter. For the 250-atom cubic unit cell using the {5,4,5;b} setting, a significant drop in total energy was seen above 5×10^{-3}. For the accuracy expected here, growth cutoffs for the first cycle were not crucial. For the {9,12,9;a} localization, we looked at several system sizes to determine the growth value

necessary to keep the same accuracy and the proper energetic ordering of the crystal phases. For the largest system we used, 10^{-3} was sufficient. The growth for the first cycle was 12,15,12 using the naming convention for the localization region. The results are presented in Table 3. The accuracy for the {5,4,5;b} localization is presented in Table 4. This localization did not obtain the proper ordering at every system size but did succeed with the proper ordering for the largest systems done.

As a test of the scaling for the 5,4,5-b setting, we present in Fig. 2 the timing for 1 GCG step per orbital vs. number of atoms. In contrast, perfect linear scaling gives a horizontal line. At smaller systems, linear scaling is not expected because the length scales are too small for the "nearsightedness" principle to take effect. The N^3 scaling of the inversion of the occupied space, $\Phi^{\dagger}S\Phi$, begins to be substantial around 1000 atoms. The timing for the 1458-atom system is smaller than expected. The impact of the N^3 parts should cause increasing deviation from a horizontal line. We attribute this discrepancy to fluctuations in timing due to running on a non-dedicated machine. For the 1458 atom system, 11.6% of the elements of $\Phi^{\dagger}S\Phi$ are non-zero so at this point an asymptotically linear scaling approach might become more efficient.

In Fig 3, we show a timing plot of diagonalization and optimization with two localization regions ({5,4,5;b} and {9,12,9;a}) vs. time (s). This displays the large difference in crossover that occurs as the localization is increased. Crossover can be obtained early, but at a cost of accuracy. These numbers are very important when deciding if a linear scaling algorithm is suitable for a certain problem. The Lapack routine DSYGVX, which calculates a given number of lowest eigenvalues and eigenvectors, from the optimized library dxml was used. All of the times are for a non-dedicated serial 440 Mhz DEC workstation. The user time from the function DTIME was used.

VII. Summary

We have introduced a method that solves for the electronic ground state in terms of localized nonorthogonal orbitals implemented in the gaussian-based density functional code, Seqquest. We have investigated the inherent length scales involved in the calculations and have altered our algorithm to be as efficient as possible for systems in the range of 200-1500 atoms. We also have discussed the geometry of the involved vector spaces with reference to recent work concerning the Grassmann manifold.

The orbital solutions can be restricted in space, a localization region, with significant savings in computational effort. In our results, we have focused on accuracy as being the motivating factor in determining how and if the new minimization method code should be used over explicit diagonalization. The accuracy of relative energies of the cubic, 2H, and 4H phases of silicon carbide have been mapped out for different localization regions. These results may benefit further investigations by providing starting points in the use of the method.

We showed increasing levels of accuracy to be obtainable with increasing spatial extent of the orbitals. This establishes the method to be a promising quantitative tool for approaches utilizing gaussians and other linear combination of atomic orbitals. The

crossover point with diagonalization for timings of the whole self-consistent cycle ranges from 200 to roughly 800 atoms depending on the accuracy desired.

Acknowledgements

Partial support was provided the Campus Laboratory collaboration of the University of California and Sandia National Laboratories. C.Y. Fong acknowledges support from a NSF grant (no. Int-9872053). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockhead Martin Company, for the United States Department of Energy, under contract No. DEAC0494AL85000.

	a (Bohr)	c/a	energy (Ry)	$E_{2(4)H} - E_c$ (mRy/atom)
Cubic	4.0950		9.7035238	
	4.1196[a]			
2H	5.7807	1.642	9.7032725	0.25139[d]
				0.06617[c]
	5.8127[a]	1.641[a]		0.19852[b]
4H	5.785	3.275	9.7035581	-0.03425[d]
				-0.139705[c]
	5.807[a]	3.271[a]		-0.08823[b]

Table 1A. Structural and energetic results with diagonalization respectively compared to experimental and theoretical values.
[a] reference 16 (Expt.)
[b] reference 17 (LMTO)
[c] reference 15 (PW)
[d] Present work (LCAO)

	Ry/atom		
# of atoms	3C	2H	4H
64	9.69864	9.70785	9.68601
128	9.69876	9.70090	9.70086
200		9.70268	9.70266
250	9.70219		
288		9.70316	9.70314
432	9.70312		
686	9.70339		
1024	9.70348	9.70326	
1176			9.70349
1458	9.70352		

Table 1B. Γ point energies derived from k-point sampling of the primitive unit cells

	Energy/atom				
	3C	ΔE(2H – 3C)	2H	ΔE(2H – 4H)	4H
Diagonalization	9.69864	0.00920	9.70785	0.02183	9.68601
{5,4,5;b}	9.69065	0.01083	9.70149	0.02328	9.67821
{4,5,5;b}	9.67891	0.01192	9.69084	0.02417	9.66667
{5,5,5;b}	9.69208	0.01171	9.70378	0.02409	9.67969
{7,5,7;b}	9.69343	0.01200	9.70543	0.02359	9.68184
{5,7,7;b}	9.69364	0.01133	9.70527	0.02336	9.68191
{7,7,7;b}	9.69418	0.01165	9.70583	0.02308	9.68275
{7,7,7;a}	9.69706	0.00983	9.70690	0.02306	9.68384
{9,7,9;a}	9.69797	0.00932	9.70729	0.02237	9.68492
{7,9,9;a}	9.69795	0.00926	9.70721	0.02216	9.68505
{9,9,9; a}	9.69824	0.00917	9.70741	0.02200	9.68541
{9,11,9;a}	9.69838	0.00918	9.70756	0.02201	9.68555
{11,9,9;a}	9.69831	0.00921	9.70752	0.02198	9.68554
{9,12,9;a}	9.69860	0.00905	9.70765	0.02190	9.68575
{12,9,9;a}	9.69843	0.00911	9.70756	0.02195	9.68561
{11,11,11;a}	9.69843	0.00921	9.70765	0.02203	9.68562
{12,12,12;a}	9.69863	0.00914	9.70777	0.02188	9.68589

Table 2. Energy in units of Ry/atom for varying localization regions in the 64 atom SiC system.

{9,12,9;a}	Energy/atom		
	3C	2H	4H
128	9.69850	9.70053	9.70048
200		9.70222	9.70218
250	9.70177		
288		9.70269	9.70263

Table 3. setting of {9,12,9;a} with growth of 12,15,12 for 1st SCF cycle and 10^{-3} for subsequent cycles

{5,4,5;a}	Energy/atom		
	3C	2H	4H
128	9.69078	9.69089	9.69110
200		9.69162	9.69183
250	9.69296		

288		9.69199	9.69209
432	9.69270		
686	9.69244		
1024	9.69349	9.69159	
1176			9.69267
1458	9.69277		

Table 4. setting of {5,4,5;a} with growth of 10,10,10 for 1st SCF cycle and $5x10^{-3}$ for subsequent cycles

1 E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10 008 (1994).
2 P. OrdejÓn, D. A. Drabold, R. M. Martin, et al., Phys. Rev. B **51**, 1456 (1995).
3 F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).
4 W. Hierse and E. B. Stechel, Phys. Rev. B **50**, 17 811 (1994).
5 J. Kim, F. Mauri, and G. Galli, Phys. Rev. B **52**, 1640 (1995).
6 U. Stephan, D. A. Drabold, and R. M. Martin, Phys. Rev. B **58**, 13 472 (1998).
7 M. P. Teter, M. C. Payne, and D. C. Allan, Phys. Rev. B **40**, 12 255 (1989).
8 M. J. Gillan, J. Phys.: Condens. Matter **1**, 689 (1989).
9 G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
10 R. D. King-Smith and D. Vanderbilt, Phys. Rev. B **49**, 5828 (1994).
11 I. Stich, R. Car, M. Parrinello, et al., Phys. Rev. B **39**, 4997 (1989).
12 X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).
13 M. S. Daw, Phys. Rev. B **47**, 10895 (1993).
14 R. W. Nunes and D. Vanderbilt, Phys. Rev. B **50**, 17 611 (1994).
15 D. R. Bowler and M. J. Gillan, Comp. Phys. Com. **120**, 95 (1999).
16 A. H. R. Palser and D. E. Manolopoulos, Phys. Rev. B. **58**, 12 704 (1998).
17 J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997).
18 M. Challacombe, J. Chem. Phys. **110**, 2332 (1999).
19 W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
20 A. Edelman, T. A. Arias, and S. T. Smith, SIAM J. on Matrix Anal. Appl. **20**, 303 (1998).
21 R. A. Lippert and M. P. Sears, Report SAND99-2986, Sandia National Laboratories, Abuquerque, NM USA, (1999).
22 A. Edelman and S. T. Smith, BIT **36**, 494 (1996).
23 P.-O. LÖwdin, Int. J. Quantum Chem. **2**, 867 (1968).
24 R. Mcweeny, Rev. Mod. Phys. **32**, 335 (1960).
25 J. Pask, B. M. Klein, C. Y. Fong, et al., Phys. Rev. B **59**, 12 352 (1999).
26 E. Tsuchida and M. Tsukada, J. Phys. Soc. Japan **67**, 3844 (1998).
27 P. Schultz and P. J. Feibelman, to be published.
28 C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
29 K. N. Kudin and G. E. Scuseria, Chem. Phys. Lett. **289**, 611 (1998).
30 E. Schwegler, M. Challacombe, and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997).
31 R. E. Stratmann, G. E. Scuseria, and M. J. Frisch, Chem. Phys. Lett. **257**, 213 (1996).
32 J. M. Perez-Jorda and W. Yang, Chem. Phys. Lett. **241**, 469 (1995).
33 O. Sinanoglu, Theoret. Chim. Acta **65**, 233 (1984).
34 E. Artacho and L. M. d. Bosch, Phys. Rev. A **43**, 5770 (1991).
35 E. B. Stechel, in *Topics in Computational Materials Science*, edited by C. Y. Fong (World Scientific, 1998), p. 1.
36 B. F. Schutz, *Geometrical methods of mathematical physics* (Cambridge University Press, 1980).
37 R. W. R. Darling, *Differential Forms and Connections* (Cambridge University Press, 1994).
38 S. Liu, J. M. Perez-Jorda, and W. Yang, J. Chem. Phys. **112**, 1634 (2000).
39 P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
40 R. A. Lippert and M. P. Sears, submitted to Phys. Rev. B, (2000).
41 C. A. White, P. Maslen, M. S. Lee, et al., Chem. Phys. Lett. **276**, 133 (1997).

[42] E. L. Briggs, D. J. Sullivan, and J. Bernholc, Phys. Rev. B 54, 14 362 (1996).
[43] J. R. Chelikowsky, N. Troullier, and Y. Saad, Phys. Rev. Lett. 72, 1240 (1994).
[44] N. A. Modine, G. Zumbach, and E. Kaxiras, Phys. Rev. B 55, 10 289 (1997).
[45] V. A. Gubanov and C. Y. Fong, Appl. Phys Lett. 75, 88 (1999).
[46] P. Käckell, B. Wenzien, and F. Bechstedt, Phys. Rev. B 50, 17 037 (1994).
[47] C. H. Park, B.-H. Cheong, K.-H. Lee, et al., Phys. Rev. B 49, 4485 (1994).
[48] S. Limpijumnong and W. R. L. Lambrecht, Phys. Rev. B 57, 12 017 (1998).
[49] H. J. Monkhorst and J. D. Pack, Phys. Rev. B 13, 5188 (1976).

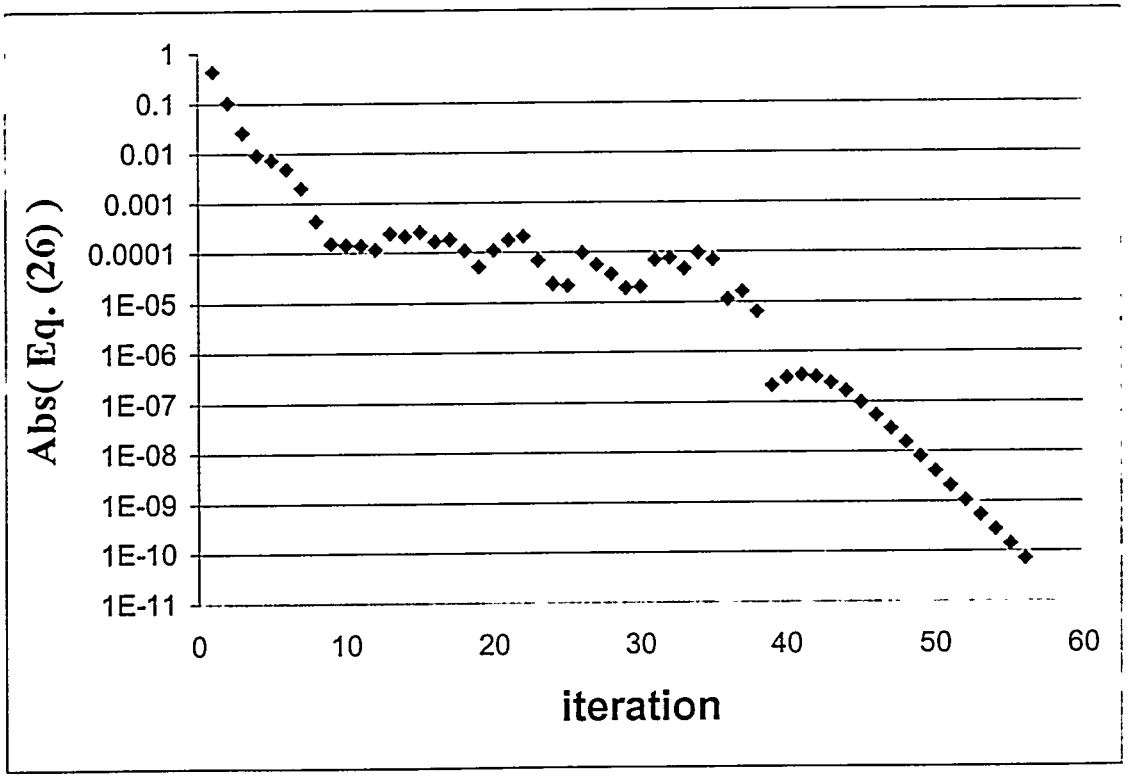

Fig. 1 G • $dE_B/d\Phi$ vs. iteration number for the 5,4,5 – b setting for the 64 atom unit cell of SiC.

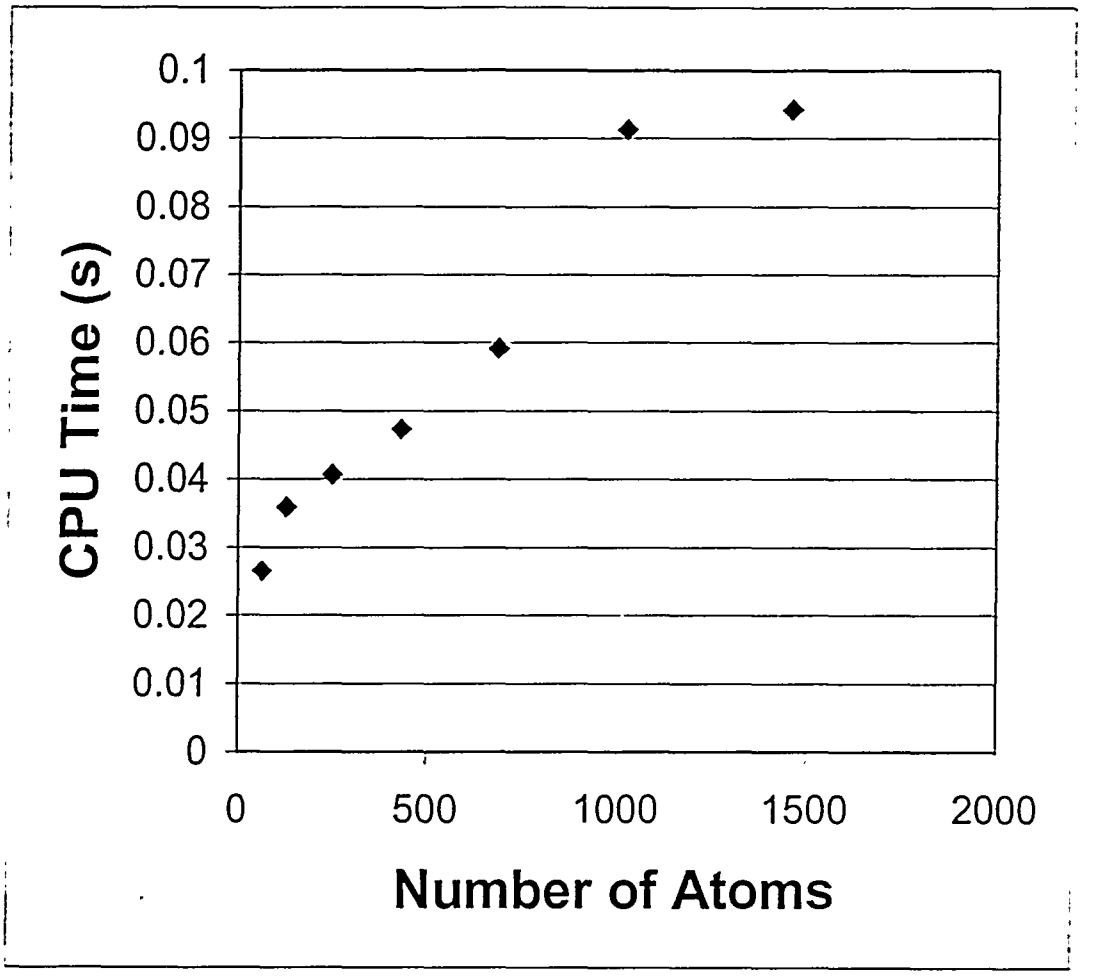

Fig 2. CPU time for one GCG step per orbital for the 5,4,5-b setting Vs. # of atoms.