5-2007

Self-consistent and environment-dependent Hamiltonian for quantum-mechanics materials simulations.

Christopher R. Leahy 1972-
University of Louisville

SELF-CONSISTENT AND ENVIRONMENT-DEPENDENT HAMILTONIAN FOR QUANTUM-MECHANICS MATERIALS SIMULATIONS

By

Chris Leahy
M.S., University of Louisville, 1998

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Physics
University of Louisville
Louisville, Kentucky

May 2007

SELF-CONSISTENT AND ENVIRONMENT-DEPENDENT HAMILTONIAN FOR QUANTUM-MECHANICS MATERIALS SIMULATIONS

By

Chris Leahy
M.S., University of Louisville, 1998

A Dissertation Approved On

_____
Date

by the following Dissertation Committee:

_____
Dissertation Director

_____

_____

_____

_____

ACKNOWLEDGEMENTS

I would like to thank the committee members, Professors Buchanan, DuPré, Jayanthi, Pack, and Wu, for reading my thesis and for their guidance during my time as a student in this program.

I would also like to express my gratitude to Professors Wu and Jayanthi for suggesting this research and for their guidance and encouragement throughout this project.

I would also like to thank my colleagues, I. Chowdhury, C. Ghosh, S. Shen, H. Simrall, L. Smith, P. Tandy, M. Yu, A. Tchernatinsky, in the Condensed Matter Theory Group for their support, particularly Drs. M. Yu and A. Tchernatinsky, and Mr. S. Shen, for their support and assistance.

I would also like to acknowledge funding for this research from a University Fellowship, and from the National Science Foundation and the U.S. Department of Energy.

ABSTRACT

SELF-CONSISTENT AND ENVIRONMENT-DEPENDENT HAMILTONIAN FOR QUANTUM-MECHANICS MATERIALS SIMULATIONS

Chris Leahy

May 12, 2007

I will report the development of a semi-empirical self-consistent and environment-dependent model Hamiltonian, which is intended to treat large systems in the order of 10000 atoms. This covers a range of important physical phenomena that are too large to be treated with first-principles calculations. Our model features an aggressive treatment of environment-dependent effects, which are known to limit the accuracy of two-center models which do not include them. Specifically, we account for multi-center integrals, and we use a full iterative treatment of the self-consistency problem, which addressed the important role of charge redistribution. Our results indicate that our treatment is superior to other semi-empirical models that treat environment-dependency in a more phenomenological manner, and either ignore the charge redistribution, or treat it not at equal footing as the environment-dependency. The feasibility of this methodology has been tested for silicon.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The field of computational materials science always goes back to the fact that a straightforward application of the known laws of physics to a problem of real-world interest, results in a burden of calculation far beyond the capabilities of any known calculating machine. The ever-increasing availability of faster computers at lower prices can not alleviate this problem, in part because the required calculations scale very slowly, but also because there is an ever-increasing demand for more complicated problems, resulting in something of an arms race among those involved. All is not lost, however, as even though a *straightforward* application of the laws of physics is not workable, a *complicated* application of the laws of physics is. Starting with the laws of quantum mechanics, these workable models can be interpreted as applying various layers of approximations to reduce the burden of the calculation. The goal is to cut out as many calculations as possible while still maintaining some meaningful level of accuracy.

It is useful to make some gross classifications of the wide variety of resulting models. First, one can distinguish models that use only fundamental physical constants from those that use adjustable parameters. Models in the first category are called "first-principles" or "ab-initio", while those in the second are usually called "empirical" or "semi-empirical". Next, one can distinguish models that calculate electronic structure properties from those that do not. The first category is usually recognizable by the existence of an eigenvalue problem, where particle interactions contribute to a Hamiltonian matrix. The second category includes "molecular mechanics" models, which replace the eigenvalue problem with a more

Newtonian formula where particle interactions contribute directly to the total energy. Finally, one can distinguish models that account for the locations of individual atoms from those that do not. This second category includes "finite element methods", which, although usually encountered in engineering problems, are increasingly seen in computational physics and chemistry.

Although useful, these gross classifications are becoming increasingly blurred as models become more complicated. First-principles methods can be chosen based on their accuracy in calculating experimentally known properties; in some ways this choice itself amounts to an empirical selection. At the same time, empirical models can be adjusted to match the results of first-principles calculations, and one can then argue that the resulting model is in some ways not empirical at all. Along these same lines is the increasing use of combining parts from different categories to produce hybrid and multi-scale models. Multi-scale modeling takes advantage of the fact that interesting physical processes often have a very small region where something interesting is happening, surrounded by a much larger region that serves mainly as ballast.

One should be careful to avoid arguments that, for example, empirical models are better than first principles models. First-principles models address the need for calculations on systems of very limited size (typically not more than 100 atoms), while empirical electronic structure models address the need for larger systems (10000 to 20000 atoms), and molecular mechanics can treat systems into the millions of atoms. In practice there is usually only one class of models suitable for a specific problem; a 1000 atom oxide surface calculation almost certainly means that one will be using an empirical electronic structure model.

This report concerns the development of a model in the category that uses adjustable parameters, the category that calculates electronic structure, and of course the category that accounts for the locations of individual atoms. The model is intended to address a wide range of important problems in materials science that

involve 1000s of atoms. This includes semiconductor surfaces, most notably Si, which is a more mature and arguably over-studied area of materials science. Also carbon nanotubes and related "nano-structures", which are currently the most popular applications, although the future of nanotubes as a consumer technology is not clear. Perhaps more interesting are potential applications in less saturated areas, such as oxides and transition metal surfaces.

Deserving special mention are biological applications. Although widely studied using electronic structure calculations, there appears to be a sharp divide between models such as our own, which have their origins in the semiconductor community, and those that are currently used to study biological systems. This divide might be related to the larger numbers of atoms needed for biological calculations, although it is probably due more to historical patterns of specialization in narrow areas of research. In any event, the types of models that are the subject of this report are rarely used for biological applications, which makes their potential for use in this area very interesting.

The primary goal of this research was the development of a self-consistent and environment-dependent model or methodology intended to be used to study large-scale systems. Although silicon is used as a representative example, the methodology has been developed with a broader range of materials in mind. Indeed, an important part of this thesis is the development of second "prototype" model that addresses the need to extend such models to organic and biological materials. During the course of the research, a significant number of insights into orbital models themselves were obtained, which are interesting outside the context of any specific calculation. Such results are discussed throughout this thesis along with the discussion of our environment-dependent model.

In addition to the successful development of our environment-dependent model, it is useful to point out here some of the more general conclusions that present themselves in this work:

- The two-center part of such orbital models is usually not given enough attention; indeed a careful treatment of the two-center part appears to be critical to the success of the overall model. This is discussed in detail in Chapter II.

- In addition to what might be called a "derivation" approach to developing such models, i.e. starting with a set of equations and attempting to derive approximate solutions, a "policy-based" approach, which starts with a list of requirements that a model must satisfy, appears to be very useful. The radial function prototype discussed in Chapter IV was obtained primarily from such a policy-based approach.

- The actual source code needed to perform any numerical calculation is also usually not given enough attention. Although our discussion does not emphasize this issue as much as others, our discussion of optimization algorithms in Chapter V serves as a relatively brief example of the large amount of "nuts-and-bolts" work that has gone into our model.

- Semi-empirical orbital models, despite being widely used in modern research, still carry a large amount of obsolete "baggage" from the early years of their development. This concept appears in our discussion of the limitations of existing models in Chapter III.

The thesis is organized as follows: In Chapter II, we discuss the two-center part of our model. This includes the development of a parameterized functional form for the overlap and Hamiltonian matrix elements. We also discuss some important results that can be thought of as more highly theoretical, i.e. results that appear to be valid outside the context of any particular material. This includes a novel derivation or interpretation of the widely used Hückel approximation, and also a new interpretation of orbital-based models in terms of the limiting values of the matrix elements for small values of the atomic site separation R. In Chapters III

and IV we discuss new ideas for the what might be called the "next generation" of orbital-based models. The central concept is that the current generation of models is simply not suitable for more complex calculations and materials. Specifically, this occurs when one attempts to model (1) the d orbitals, (2) multi-element systems, and (3) organic materials, specifically those involving nitrogen and oxygen. A new radial function prototype is presented which addresses these issues.

In Chapter V we turn our attention to optimization algorithms, which are a central part of the source code that is used to obtain the semi-empirical parameters. One of the purposes of this chapter is to illustrate, with selected examples, the large amount of work that was done in the development of the actual source code. Indeed, this is the area in which my own work was most heavily concentrated. Finally, in Chapter VI we discuss the environment-dependent parts of our model, i.e. the parts of our model that were not discussed in Chapter II. Also in this chapter we report the parameterization and of our model for silicon, and we show some representative applications of this model: the structure of the intermediate-sized Si_{71} cluster, and the reconstruction of the Si (001) surface.

CHAPTER II

TWO-CENTER TECHNIQUES

A "Nine functions"

The atomic-scale modeling technique that we use goes by a variety of names. In the past, it was referred to as *tight-binding*, although this name now often refers to less-computational and more analytical techniques. The name *linear combination of atomic orbitals (LCAO)* is appropriate, although this name also describes several other techniques. Since the technique uses free parameters or empirical parameters chosen to give the best calculated values, it can be referred to as *parameterized* or *empirical* or *semi-empirical*. The atomic-scale interactions are based on two-center integrals, with modifications for higher-order interactions called environment-dependent interactions. So, the names *two-center* and *environment-dependent* can also be used. Since the two-center integrals use a non-orthogonal basis set, the name *non-orthogonal* is also occasionally used. Finally, the higher-order interactions involve a self-consistent calculation of the electron numbers, and so the name *self-consistent* can also be used.

The two-center part of the technique, i.e. without the environment-dependent modifications, has a long history. The development of this technique can be interpreted as a series of layers of approximations, starting with the fundamental equation of quantum mechanics, the Schrödinger equation. This equation is intractably difficult to solve, either analytically or numerically, for any system with more than a total of a few electrons and nuclei. So, a series of approximations are made to obtain a tractable computational problem. We will discuss certain areas of

these approximations in detail as they relate to our specific empirical orbital model. However, a detailed discussion of each layer of approximation is outside the scope of this thesis.

The result of these approximations is that physical properties can be calculated from a small number of two-center interactions or *two-center integrals*. Specifically, each interaction is a scalar function of the scalar separation R between two atoms. Loosely speaking, each function represents the strength of a particular type of interaction between the atomic orbitals of two atoms separated by a distance R. For a basis set consisting of s and p orbitals, there are a total of 9 such functions, 4 each for the overlap and Hamiltonian interactions, and 1 for a two-body repulsive interaction:

$$\begin{array}{ll} \text{overlap:} & S_{ss\sigma}(R),\ S_{sp\sigma}(R),\ S_{pp\sigma}(R),\ S_{pp\pi}(R) \\ \text{Hamiltonian:} & H_{ss\sigma}(R),\ H_{sp\sigma}(R),\ H_{pp\sigma}(R),\ H_{pp\pi}(R) \\ \text{repulsive:} & E_{rep}(R) \end{array} \tag{1}$$

The early development of these functions is attributed to Slater [1]. For a particular configuration of atoms then, these overlap and Hamiltonian functions are used to set up the overlap matrix S and the Hamiltonian matrix H. The resulting eigenvalue equation is then solved for the energy eigenvalues E. The electrons are then assigned to the energy eigenvalues using a distribution such as the Fermi distribution. The resulting energy is the band energy E_{band}. The band energy is then combined with the repulsive energy E_{rep} to give the total energy E_{tot}.

Unfortunately, after the pre-computational 1954 paper by Slater, it is not clear exactly who to attribute the later development of these functions to. Harrison [2] developed much of the early less-computational "tight-binding" theory, which was widely used to obtain closed-form analytical expressions for material properties. Chadi [3], in a series of papers in the late 1970s, developed a widely-used orthogonal model. These types of early orthogonal models did not use overlap functions S or a repulsive energy E_{rep}. It is also Chadi [4] who is credited with using a two-body

repulsive interaction in 1979 (earlier models often did not require the specification of a total energy because they calculated properties that depended only on the band energy E_{band}). Tománek and Schülter [5] are usually credited with applying these types of models to clusters in 1986 (earlier models were almost exclusively for systems with periodic boundary conditions, such as crystalline Si). One of the earliest models to use overlap functions, i.e. a non-orthogonal model, is that of Allen, Broughton, and McMahan in 1986 [6]. In 1992 Wang and Ho [7] developed a model for both cluster and bulk C, and in 1993 Mercer and Chou [8] developed a similar model for Si and Ge. In our opinion these are the first two models that have the same "look and feel" as the models that are currently in use.

Now, we are using a parameterized technique, which means that the shape of each of these 9 functions will be adjusted to give the best calculated values. The implementation of such a technique into a computer program requires the development and testing of a large amount of source code, which is quite difficult and time-consuming. Still, since the calculated energies are determined entirely by the 9 scalar functions, it seems that this two-center model should be a "closed case". What remains to be said about the two-center model? Actually, a *great deal* remains to be said. Improvements to the two-center model have been a major success of our work. These include improvements to the computational model, which should be of interest to the atomic-scale modeling community, and also improvements to the theoretical interpretation of empirical orbital models, which should be of interest to the broader community. So, in this chapter we will discuss our two-center model, i.e. the two-center part of our model before we apply our environment-dependent modifications.

B Hyperbolic function

The first item that we need is a parameterized functional form for the 8 overlap and Hamiltonian functions in eq. 1. The repulsive energy is treated

separately. Now, from a less-computational perspective, one is interested in the best shape of the entire function, i.e. as if the function had an infinite number of parameters, where each parameter would be the value of the function at a specific value of R, and where R can take on all values from 0 to ∞. However, for the numerical problem one needs a relatively small number of parameters. A brute-force attempt might be to use a grid or mesh of around 200 points uniformly spaced from $R = 0$ to some maximum value $R = R_{max}$. This would result in around 1600 empirical parameters, i.e. 200 for each of the 8 functions in eq. 1, which is quite beyond the capabilities of a modern computer system. A second attempt might be to use a function such as a polynomial, and to treat the coefficients of the polynomial as parameters. To allow for a function of a reasonably arbitrary shape, i.e. a smooth function without too many oscillations, it would be necessary to use about 10 or 12 coefficients or parameters for each function. This would result in around 100 empirical parameters. Now, if finding the best set of parameters is a *local* optimization problem, then one can probably use 100 parameters. However, we have found after much experimentation that finding the best set of parameters is a *global* optimization problem. We have also found that this is a particularly difficult global optimization problem. It is our position that, with the computational resources currently available, it is not possible to find the global minimum with reasonable confidence for such a large number of parameters.

Our search for a functional form with a smaller number of parameters began by considering the work of Frauenheim et. al. in Ref. [9], [10], [11]. Frauenheim used first-principles calculations to obtain the two-center integrals in eq. 1 for Carbon, Silicon, and Germanium. Their method, described as "density functional tight binding", consists of solving a modified version of the atomic Kohn-Sham equations for each element of interest. The eigenfunctions obtained from the Kohn-Sham equations are then used to construct the two-center integrals. It is important to clarify that we do not expect our final parameters, i.e. after empirical

fitting, to exactly reproduce these first-principles integrals. Loosely speaking, these functions serve as "internal" quantities that one does not expect to be able to compare with any experiment. Different first-principles methods will give different values for these integrals. Even if they did all give the same values, the concept of empirical modeling is to allow certain quantities that are not of interest to the "end user" to have values that are slightly different from the known values. This flexibility allows other quantities that are of interest to the end user to have values that are more accurate, i.e. the empirical model recovers some of the accuracy that is lost in the various layers of approximations discussed in Section A.

The integrals of Frauenheim are shown in Figure 1 for Silicon. We noted that for each function there appear to be two different regions of behavior; the first for $R > 2.0$Å where the functions are quickly decreasing to zero, and the second for $R < 2.0$Å where the shape is linear. The behavior for $R > 2.0$Å is due to the physical constraint that the interactions must go to zero outside a small range. The behavior for $R < 2.0$Å is due to physical constraints on the integrals for $R \to 0$. The behavior for $R \approx 2.0$Å is due to a competition between these constraints. Our parameterized form begins with the concept of two regions of behavior separated by a crossover separation S_{cross}, which is treated as a free parameter. We start with a variant of the Fermi distribution function, which features such a two-region behavior:

$$S(R) \overset{\text{first attempt}}{=} \frac{\exp(-S_{exp} \cdot (R - S_{cross}))}{1 + \exp(-S_{exp} \cdot (R - S_{cross}))}$$

This form also already features the desired exponential decrease to zero; the range of the interaction is determined by the free parameter S_{exp}. To allow for a small number of oscillations in each function, we include a polynomial factor with free parameters S_0 and S_1:

$$S(R) \overset{\text{second attempt}}{=} (S_0 + S_1 \cdot R) \cdot \frac{\exp(-S_{exp} \cdot (R - S_{cross}))}{1 + \exp(-S_{exp} \cdot (R - S_{cross}))}$$

Finally, we are also interested in the value of the functions for $R \to 0$. With a small

modification we can force S_0 to be the value of the function at $R = 0$:

$$S(R) \overset{\text{final form}}{=} (S_0 + S_1 \cdot R + (S_0 + S_1 \cdot R) \cdot \exp(-S_{exp} \cdot S_{cross})) \cdot \frac{\exp(-S_{exp} \cdot (R - S_{cross}))}{1 + \exp(-S_{exp} \cdot (R - S_{cross}))} \tag{2}$$

This is the final form of our parameterized function.

As an initial test, we fit our functional form in eq. 2 to the integrals of Frauenheim for Silicon shown in Figure 1. With 8 functions, and 4 free parameters S_0, S_1, S_{exp}, S_{cross} for each function, we used a total of 32 parameters in this test. The results are shown in Figure 2. The agreement is quite remarkable; all features of the integrals are accurately reproduced. Also, the numerical values of the free parameters are in agreement with the physics of the material. For example, the values of S_{cross} are all around 2.0Å, which is the value that separates the two regions of behavior for Silicon. Similar agreement is also found for C and Ge. We should again point out that this test fitting is not in any way an attempt to obtain a final set of parameters for Si. Here, we are only demonstrating the ability of our functional form to take on the *variety of shapes* that are expected for two-center integrals.

Using 32 parameters is still a rather large number for a poorly-behaved global problem. Since our parameters can be directly related to the physics of the material, it is possible to further reduce the number of parameters using constraints. After a significant amount of experimentation, we have found that the parameters S_{exp} and S_{cross} can be constrained to have the same value for each of the 4 overlap functions. This also works for the Hamiltonian functions, i.e. with values H_{exp} and H_{cross} that are different from the overlap values. This reduces the number of parameters from 32 to 20, which is a significant improvement. However, further experimentation has shown that these constraints might not be appropriate for C, and we are also moving away from using these constraints for Si and Ge.

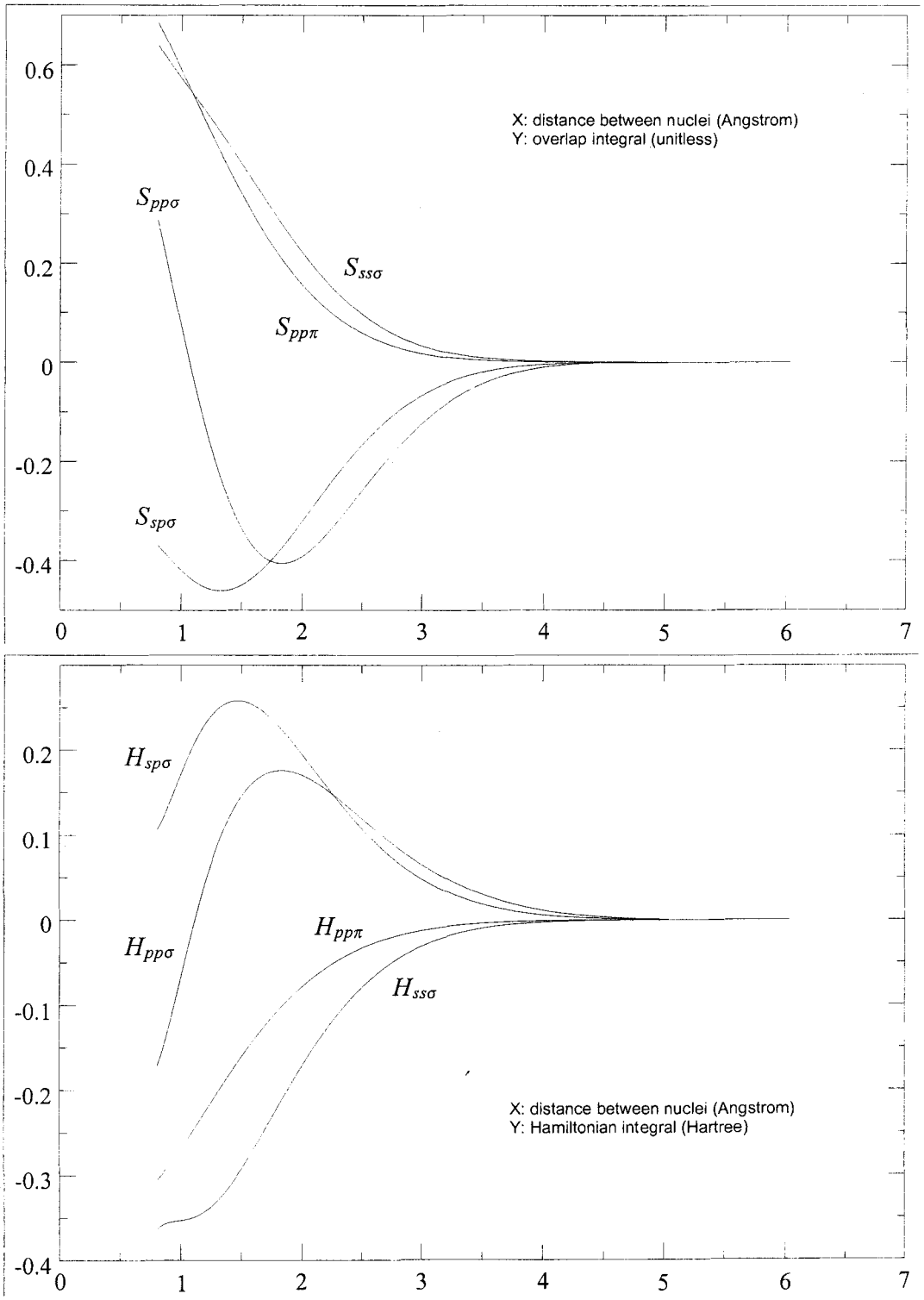

Figure 1. Overlap (top) and Hamiltonian (bottom) integrals for Si calculated using the first-principles technique of Frauenheim from Ref. [10].

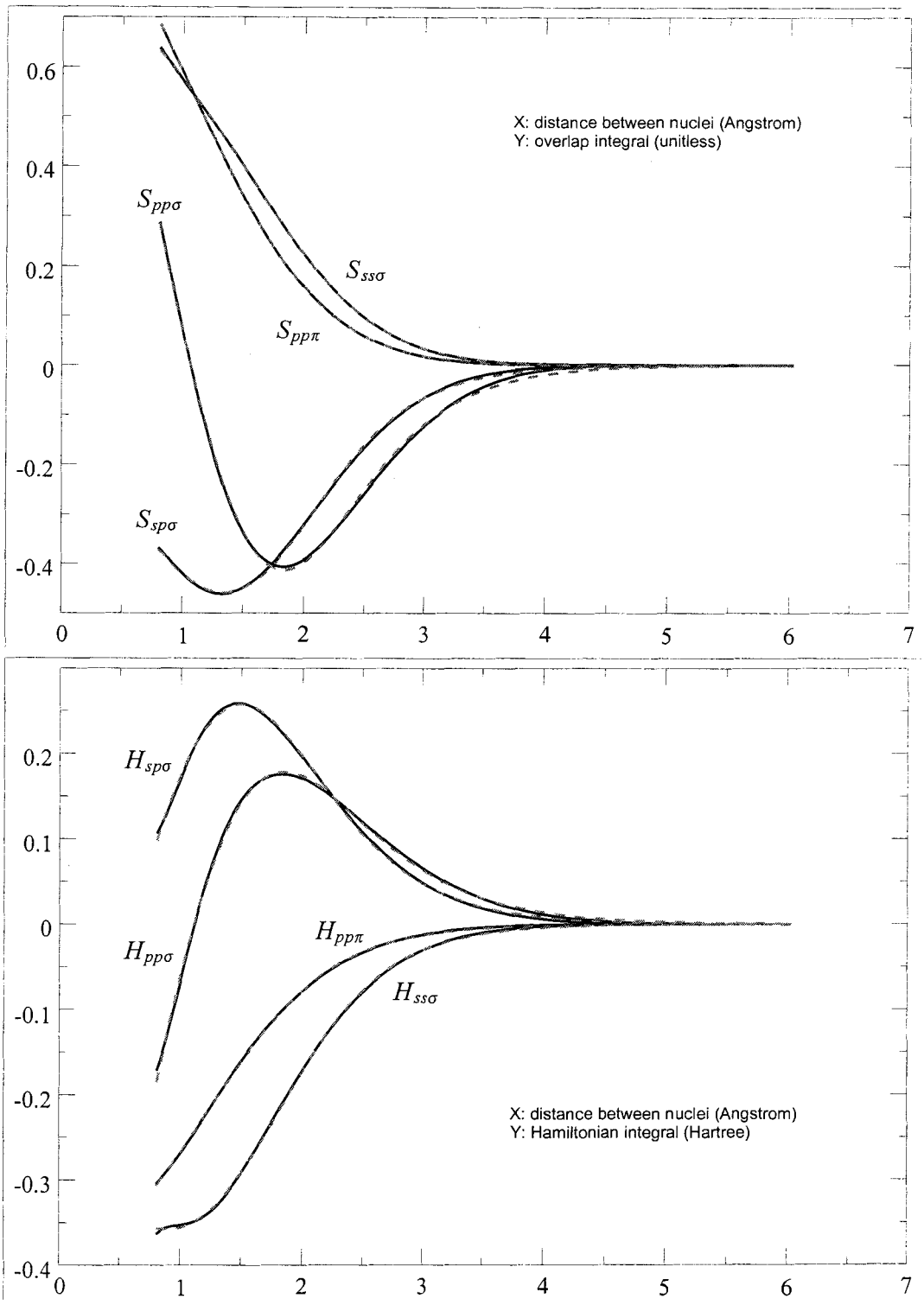

Figure 2. Comparison of overlap (top) and Hamiltonian (bottom) integrals for Si calculated using our hyperbolic function (dashed) to the first-principles integrals of Frauenheim (solid) from Ref. [10].

C Overlap integrals for $R \to 0$

For real materials, even under the most extreme conditions, atoms do not come much closer to each other than they do under normal conditions. For example, the nearest-neighbor distance for Si in a diamond anvil cell at an extreme pressure of 250GPa is only about eight percent smaller than the nearest-neighbor distance at atmospheric pressure [12]. For these reasons it is often argued that the two-center integrals in eq. 1 do not contain any useful information for small values of R. However, the strong repulsion of the atomic nuclei, which is responsible for the atoms not coming close to each other, is not present in the two-center integrals. In fact, the two-center integrals are well-defined for all values of R, including the limit as $R \to 0$, and including the values at $R = 0$. This can be seen from the explicit form of the two-center integrals:

$$S_{ss\sigma}(R) = \int_{\mathbf{r}} \Phi_s(\mathbf{r} - \mathbf{R}_1) \cdot \Phi_s(\mathbf{r} - \mathbf{R}_2) \cdot d\mathbf{r} \qquad \text{with } R = \|\mathbf{R}_2 - \mathbf{R}_1\|$$

$$H_{ss\sigma}(R) = \int_{\mathbf{r}} \Phi_s(\mathbf{r} - \mathbf{R}_1) \cdot \hat{\mathrm{H}} \cdot \Phi_s(\mathbf{r} - \mathbf{R}_2) \cdot d\mathbf{r} \qquad \text{with } R = \|\mathbf{R}_2 - \mathbf{R}_1\|$$

Here, we have used the $ss\sigma$ interaction as an example; the discussion in this section applies in general to all such two-center integrals, including other orbitals such as the d-orbitals.

Following the diamond anvil cell argument, we can see that small values of R and values in the limit as $R \to 0$ will never be present in any matrix elements for any physical system that might be of interest in the field of materials science. However, the values at $R = 0$ are *always* present; these are just the "on-site" matrix elements, which are associated with the interaction of an atomic orbital with itself and with other atomic orbitals at the same nucleus. In the literature the on-site matrix elements are often discussed without reference to R, as in:

$$S_{ss\sigma}|_{\text{on-site}} = \int_{\mathbf{r}} \Phi_s(\mathbf{r}) \cdot \Phi_s(\mathbf{r}) \cdot d\mathbf{r}$$

$$H_{ss\sigma}|_{\text{on-site}} = \int_{\mathbf{r}} \Phi_s(\mathbf{r}) \cdot \hat{\mathrm{H}} \cdot \Phi_s(\mathbf{r}) \cdot d\mathbf{r}$$

The crux of the matter is whether the diamond anvil cell argument means that one can ignore the behavior of the two-center integrals for small values of R. This has been the traditional argument, that one can use a function that has the appropriate behavior for experimentally relevant values of R, but that might be divergent or undefined for small values of R, and that the on-site matrix elements can be treated separately, without reference to R, as in this expression. We will argue in this study that the values of the two-center integrals at small values of R contain significant information that can improve both the computational aspect and also the theoretical interpretation of the model.

We will consider the overlap integrals first; these integrals are simpler since they do not involve the Hamiltonian operator $\hat{\mathrm{H}}$. If we use the analogy of a knob that can be used to turn down the value of R, we have two important results. First, the value of the integral in the limit as $R \rightarrow 0$ is equivalent to the value of the integral at $R = 0$. Second, at $R = 0$ the two-center integral becomes a one-center integral, and the value of the integral is then determined by the fact that atomic orbitals at the same site are orthogonal and normalized. Since the $ss\sigma$ interaction involves the same orbitals Φ_s and Φ_s we must have $S_{ss\sigma}(R)|_{R=0} = 1$. Similarly, since the $sp\sigma$ interaction involves two different orbitals Φ_s and Φ_p we must have $S_{sp\sigma}(R)|_{R=0} = 0$. We then have:

$$\begin{aligned} \lim_{R\rightarrow 0} S_{ss\sigma}(R) &= S_{ss\sigma}(R)|_{R=0} = 1 \\ \lim_{R\rightarrow 0} S_{sp\sigma}(R) &= S_{sp\sigma}(R)|_{R=0} = 0 \\ \lim_{R\rightarrow 0} S_{pp\sigma}(R) &= S_{pp\sigma}(R)|_{R=0} = 1 \\ \lim_{R\rightarrow 0} S_{pp\pi}(R) &= S_{pp\pi}(R)|_{R=0} = 1 \end{aligned} \tag{3}$$

One should not dismiss as trivial the result that the limit as $R \rightarrow 0$ is equivalent to the value at $R = 0$. Although this result is valid for the overlap integrals, it is not valid for the Hamiltonian integrals.

We can see clear evidence in Figure 1 and 2 that the overlap integrals extrapolate to these limiting values. This same behavior is also observed for C in

Ref. [9] and Ge in Ref. [11]. From this we can begin to make the argument that the values of the two-center integrals, at values of R that are not experimentally relevant, affect the values of the two-center integrals at values of R that *are* experimentally relevant. For example, the parameters S_0 in eq. 2, when fit to the first-principles integrals of Frauenheim, all have values around either 1.0 or 0.0 consistent with eq. 3. One could continue to develop this argument based on other features in Figure 1 and 2. For example, in the range of chemical bonding, $S_{pp\sigma}$ is always significantly larger (in magnitude) than $S_{sp\sigma}$, and this seems to be related to the different limiting values of $S_{pp\sigma}$ and $S_{sp\sigma}$. However, we feel that the best argument for the importance of treating small values of R is the practical benefit this provides to the search for the best set of parameters. We will discuss this issue in more detail, in the more general context of parameter constraints, in Section F.

D Hamiltonian integrals for $R \rightarrow 0$

The behavior of the Hamiltonian integrals for $R \rightarrow 0$ is more complicated because of the Hamiltonian operator $\hat{\mathrm{H}}$. We first need the expanded form of the Hamiltonian:

$$\hat{\mathrm{H}} = \nabla_{\mathbf{r}}^2 + V(\mathbf{r})$$

This form is valid within the mean field approximation, which is one of the layers of approximation discussed in Section A. The expanded form of the potential $V(\mathbf{r})$ is:

$$V(\mathbf{r}) = \sum_k V_k(\|\mathbf{r} - \mathbf{R}_k\|)$$

where the sum is over all the atomic nuclei indexed by k. This form is valid within the central field approximation, which is also one of the layers of approximation discussed in Section A. This will result in the following terms in the Hamiltonian matrix elements:

$$H_{ij,\nabla} = \int_{\mathbf{r}} \Phi_i(\mathbf{r} - \mathbf{R}_i) \cdot \nabla_{\mathbf{r}}^2 \cdot \Phi_j(\mathbf{r} - \mathbf{R}_j) \cdot d\mathbf{r}$$

$$H_{ij,k} = \int_{\mathbf{r}} \Phi_i \left(\mathbf{r} - \mathbf{R}_i\right) \cdot V_k \left(\|\mathbf{r} - \mathbf{R}_k\|\right) \cdot \Phi_j \left(\mathbf{r} - \mathbf{R}_j\right) \cdot d\mathbf{r} \tag{4}$$

The terms $H_{ij,k}$ are in general three-center integrals, i.e. involving the three centers or three atomic nuclei at $\mathbf{R}_i$, $\mathbf{R}_j$, $\mathbf{R}_k$. Within the two-center approximation, integrals involving three distinct centers are taken to be zero:

$$H_{ij,k} = 0 \qquad \text{for } i \neq j \neq k$$

Before proceeding, it is important to note that the layers of the mean field approximation, central field approximation, and two-center approximation are *already required* by the empirical orbital model, i.e. required in order to construct an overlap and Hamiltonian matrix from the functions in eq. 1. This means that this expansion of the Hamiltonian is exact within the approximations that are already required by the model.

We now consider the limiting behavior of the terms eq. 4 for $R \to 0$. It is now very important to carefully distinguish between the $R \to 0$ limit and the values at $R = 0$. The limiting behavior of the ∇^2 terms is not problematic; we have:

$$\lim_{R \to 0} H_{ij,\nabla}(R) = H_{ij,\nabla}(R)|_{R=0} = H_{ii,\nabla}$$

For the $H_{ij,k}$ terms, one must consider that even for very small values of R, the centers $\mathbf{R}_i$ and $\mathbf{R}_j$ are still distinct, and thus the potential terms V_i and V_j are also distinct:

$$\lim_{R \to 0} H_{ij}(R) = \lim_{R \to 0} H_{ij,\nabla} + \lim_{R \to 0} H_{ij,i} + \lim_{R \to 0} H_{ij,j} = H_{ii,\nabla} + H_{ii,i} + H_{ii,i}$$

However, for the values at $R = 0$, the centers $\mathbf{R}_i$ and $\mathbf{R}_j$ are equivalent, and thus the potential terms V_i and V_j are equivalent, i.e. there is *only one* potential term for which $k = i$ or $k = j$. Furthermore, at $R = 0$ the terms in eq. 4 that one might be tempted to discard as three-center integrals, are actually *two-center* integrals. That is, at $R = 0$ we have $i = j$, and there are no combinations of i, j, k for which $i \neq j \neq k$. There are no three-center integrals that can be discarded at $R = 0$. This

gives:

$$H_{ij}(R)|_{R=0} = H_{ii,\nabla} + H_{ii,i} + \sum_{k\neq i} H_{ii,k}$$

These results can be combined to give:

$$H_{ij}(R)|_{R=0} = \lim_{R\to 0} H_{ij}(R) - H_{ii,i} + \sum_{k\neq i} H_{ii,k} \tag{5}$$

This is a very important result that we will return to throughout this chapter.

For the purposes of our discussion in this section, eq. 5 shows that the limiting behavior of the Hamiltonian integrals is not equivalent to the values at $R = 0$. This result applies both to the more general form $H_{ij}(R)$ in eq. 5, and also to the specific forms $H_{ss\sigma}(R)$ etc. in eq. 1. For example, if we keep in mind that $H_{ss\sigma}(R)$ is just a particular type of integral $H_{ij}(R)$ where the orbitals Φ_i and Φ_j are both required to be s-orbitals Φ_s, we have:

$$\begin{aligned} H_{ss\sigma}(R)|_{R=0} = \lim_{R\to 0} H_{ss\sigma}(R) - \int_{\mathbf{r}} \Phi_s(\mathbf{r}-\mathbf{R}_i)\cdot V_i(\|\mathbf{r}-\mathbf{R}_i\|)\cdot\Phi_s(\mathbf{r}-\mathbf{R}_i)\cdot d\mathbf{r} \\ + \sum_{k\neq i}\int_{\mathbf{r}} \Phi_s(\mathbf{r}-\mathbf{R}_i)\cdot V_k(\|\mathbf{r}-\mathbf{R}_k\|)\cdot\Phi_s(\mathbf{r}-\mathbf{R}_i)\cdot d\mathbf{r} \end{aligned}$$

We can now proceed to set up the of constraints on the Hamiltonian integrals as we did on the overlap integrals in eq. 3. For this we will need two additional results:

$$\lim_{R\to 0} H_{sp\sigma}(R) = 0$$

$$\lim_{R\to 0} H_{pp\pi}(R) = \lim_{R\to 0} H_{pp\sigma}(R)$$

These results can be obtained by noting that both the operator $\nabla^2_{\mathbf{r}}$ and, within the two-center approximation, the potential terms $V_i(\|\mathbf{r}-\mathbf{R}_i\|)$ and $V_j(\|\mathbf{r}-\mathbf{R}_j\|)$, modify only the radial parts of the orbitals $\Phi(\mathbf{r}-\mathbf{R})$ and not the angular parts. As $R\to 0$, the angular parts of the integrals become orthogonal, but the modified radial parts do not become normalized. This gives:

$$\begin{aligned} \lim_{R\to 0} H_{ss\sigma}(R) &= \varepsilon'_s \neq H_{ss\sigma}(R)|_{R=0} \\ \lim_{R\to 0} H_{sp\sigma}(R) &= 0 \neq H_{sp\sigma}(R)|_{R=0} \\ \lim_{R\to 0} H_{pp\sigma}(R) &= \varepsilon'_p \neq H_{pp\sigma}(R)|_{R=0} \\ \lim_{R\to 0} H_{pp\pi}(R) &= \varepsilon'_p \neq H_{pp\pi}(R)|_{R=0} \end{aligned} \tag{6}$$

which can be interpreted as a definition of ε'_s and ε'_p. Here, we have used ε'_s and ε'_p rather than ε_s and ε_p; these latter symbols are usually already reserved for the on-site energies, which will be discussed later.

E Hamiltonian integrals at $R = 0$

In Section C and D we considered the relationship between the $R \to 0$ limit and the values at $R = 0$, with the goal of obtaining expressions for the $R \to 0$ limit in eq. 3 and 6. In this section we consider the same relationship, but with the goal of obtaining expressions for the values at $R = 0$. Historically, the values at $R = 0$, called the on-site energies, were developed first. However, following the order of the development in this chapter, we will take the perspective that we already have a set of 8 functions $H_{ss\sigma}(R)$ etc. with well-defined $R \to 0$ limits, and that we still need to specify how our empirical orbital model is going to treat the functions at $R = 0$. For the overlap functions, we already have the values at $R = 0$ from eq. 3, and so it remains only to treat the Hamiltonian functions. Of course, if one wants to take the more historical perspective that one already has the on-site energies and that one wants to specify the $R \to 0$ limit in terms of these on-site energies, it is a straightforward matter to work backwards from the results in this section to the results in Section C and D.

Let us then return to eq. 5:

$$H_{ij}(R)|_{R=0} = \lim_{R\to 0} H_{ij}(R) - H_{ii,i} + \sum_{k\neq i} H_{ii,k}$$

Following a more historical perspective, one can argue that $H_{ii,i}$ is expected to have a larger contribution to the Hamiltonian than $\sum_{k\neq i} H_{ii,k}$ because $H_{ii,i}$ is a one-center integral while all the terms in $\sum_{k\neq i} H_{ii,k}$ are two-center integrals. An alternative argument that leads to the same result is that $\sum_{k\neq i} H_{ii,k}$ must be discarded in order to treat the values at $R = 0$ as empirical parameters. This is due to the fact that the value of $\sum_{k\neq i} H_{ii,k}$ depends on a specific configuration of atoms, i.e. depends on

the all the potentials $V_k\left(\|\mathbf{r}-\mathbf{R}_k\|\right)$ located at all the other atomic nuclei at $\mathbf{R}_k$ (see eq. 4). Following the same arguments as in Section D, we can then show that:

$$H_{sp\sigma}(R)|_{R=0} \overset{\text{not quite}}{=} 0$$

$$H_{pp\pi}(R)|_{R=0} \overset{\text{not quite}}{=} H_{pp\sigma}(R)|_{R=0}$$

and:

$$\begin{aligned} H_{ss\sigma}(R)|_{R=0} &\overset{\text{not quite}}{=} \varepsilon_s \\ H_{sp\sigma}(R)|_{R=0} &\overset{\text{not quite}}{=} 0 \\ H_{pp\sigma}(R)|_{R=0} &\overset{\text{not quite}}{=} \varepsilon_p \\ H_{pp\pi}(R)|_{R=0} &\overset{\text{not quite}}{=} \varepsilon_p \end{aligned} \tag{7}$$

which can be interpreted as a definition of ε_s and ε_p. These expressions are valid only if all the terms in $\sum_{k\neq i} H_{ii,k}$ are neglected.

For the Hamiltonian matrix then, one has 4 functions of R with well-defined limits ε' for $R \to 0$, and with well-defined values ε at $R = 0$, but with $\varepsilon' \neq \varepsilon$. Before continuing, it is useful to look ahead to some of the results that will be developed later in this chapter by considering the expected numerical values for the energies ε and ε'. Since we are dealing with bound states, the on-site energies ε are expected to be negative. We can go so far as to obtain explicit values for the on-site energies by considering that, for a system consisting of only a single atom, the energy eigenvalues are just the on-site energies ε. For example, a density-functional theory calculation for a single atom gives the following results:

	C	Si	Ge
ε_s	-0.563 Ht	-0.439 Ht	-0.463 Ht
ε_p	-0.223 Ht	-0.166 Ht	-0.161 Ht

(8)

These results were obtained using the Gaussian-03 software package with the MPW1PW91 hybrid functional and the cc-pVTZ basis set. As with any DFT calculation, there are several caveats about how to obtain and interpret the results; however, we are using the results here only as a representative example.

We can now return to eq. 5, substituting the results in eq. 7, 6, and 4 to obtain:

$$\varepsilon_s = \varepsilon'_s - \int_{\mathbf{r}} \Phi_s(\mathbf{r} - \mathbf{R}_i) \cdot V_i(\|\mathbf{r} - \mathbf{R}_i\|) \cdot \Phi_s(\mathbf{r} - \mathbf{R}_i) \cdot d\mathbf{r}$$
$$\varepsilon_p = \varepsilon'_p - \int_{\mathbf{r}} \Phi_p(\mathbf{r} - \mathbf{R}_i) \cdot V_i(\|\mathbf{r} - \mathbf{R}_i\|) \cdot \Phi_p(\mathbf{r} - \mathbf{R}_i) \cdot d\mathbf{r}$$

Since we are still dealing with bound states, the integrals in this expression, which involve the potential V, are expected to be negative. This means that the limiting values ε'_s and ε'_p are expected to be *more negative* than the on-site values ε_s and ε_p. We can obtain approximate numerical values for ε'_s and ε'_p by noting that the on-site energies ε_s and ε_p and the integrals in this expression are both strictly one-center integrals. In the absence of any other information, one might expect the values of the potential integrals to be approximately equal to the values of the on-site energies, giving:

$$\begin{aligned} \varepsilon'_s &\sim 2 \cdot \varepsilon_s \\ \varepsilon'_p &\sim 2 \cdot \varepsilon_p \end{aligned} \tag{9}$$

If one considers only the fit in Figure 2, the evidence for this limiting behavior of the Hamiltonian is inconclusive. Indeed, one can creatively extrapolate the functions in Figure 2 to just about any energy from 0.0Ht to −1.0Ht. One issue here is that the MPW1PW91 energies in eq. 8 are not the on-site energies used by Frauenheim. It is also quite possible that the functions in Figure 2 are showing the limitations of the mean field approximation and the central field approximation, which are needed to obtain eq. 6. However, as with the overlap functions, we feel that the evidence for this limiting behavior is provided by our extensive experience in obtaining reasonable results by incorporating this limiting behavior in our model, and our extensive experience in obtaining *unreasonable* results without this. For example, we have found that, when incorporated into a full-scale empirical optimization, we can not obtain reasonable results with $\varepsilon' = \varepsilon$. This invariably leads to a poor fit, or to parameter values which are not physically meaningful.

F Parameter constraints

In Section B, we discussed that the parameters S_{exp} and S_{cross} can be constrained to have the same value for each of the 4 overlap functions, and that the parameters H_{exp} and H_{cross} can be constrained to have the same value for each of the 4 Hamiltonian functions. We also showed that for each function, the parameter S_0 or H_0 is the value of the function in the limit as $R \to 0$. From our discussion of the behavior of these functions for small values of R, we have obtained expressions for the values of these functions in the $R \to 0$ limit and for the values at $R = 0$ in eq. 3, 6, and 7. These results can be used to further reduce the number of empirical parameters. One can then select from several different models or parameterizations depending on which constraints are used. First, there is the most general parameterization with no constraints:

model with no parameter constraints:

4 parameters for each overlap function

4 parameters for each Hamiltonian function

on-site energies ε_s and ε_p

34 total parameters (two-center part only)

Then there is a parameterization with the *exp* and *cross* constraints discussed in Section B:

model with *exp* and *cross* constraints:

2 parameters for each overlap function, plus S_{exp} and S_{cross}

2 parameters for each Hamiltonian function, plus H_{exp} and H_{cross}

on-site energies ε_s and ε_p

22 total parameters (two-center part only)

Then there is a parameterization with the $R \rightarrow 0$ constraints. Here, the parameters S_0 or H_0 are replaced with the appropriate limiting values from eq. 3 and 6:

model with $R \rightarrow 0$ constraints:

---

3 parameters for each overlap function

3 parameters for each Hamiltonian function, plus ε'_s and ε'_p

on-site energies ε_s and ε_p

28 total parameters (two-center part only)

Finally, there is a parameterization with both the *exp* and *cross* constraints and the $R \rightarrow 0$ constraints:

model with *exp* and *cross* and $R \rightarrow 0$ constraints:

---

1 parameter for each overlap function, plus S_{exp} and S_{cross}

1 parameter for each Hamiltonian function, plus H_{exp} and H_{cross}, plus ε'_s and ε'_p

on-site energies ε_s and ε_p

16 total parameters (two-center part only)

(10)

These parameterizations address two key issues in empirical modeling: to *reduce* the total number of parameters, and to provide accurate *initial values* for the parameters. To understand the importance of these issues, we must keep in mind that our empirical optimization problem requires a very large number of evaluations of a least-squares function, i.e. a number large enough to exhaust any computational resources that we might have available. For such problems, if one uses too many parameters, or if one uses inaccurate initial values for the parameters, it is quite possible to have an unsolvable or intractable numerical problem. By applying parameter constraints, we have reduced the number of parameters to about 20 for the two-center integrals (the final total number of parameters will be about 2× this with the environment-dependent modifications, which are not discussed in this chapter).

These constraints also allow us to provide accurate initial values for the

parameters. That is, even if the constraints are not applied *during* the fitting process, they can still be applied *before* the fitting process to obtain the initial values. Values for the on-site energies ε_s and ε_p can be obtained directly from first-principles calculations as in eq. 8. Values for ε'_s and ε'_p can be obtained from eq. 1. Values for S_{exp}, S_{cross}, H_{exp}, H_{cross} can be obtained from a knowledge of the two-atom cluster of the element of interest. In the most straightforward case, S_{cross} and H_{cross} are just the equilibrium dimer bond length or nearest-neighbor distance of the material, and S_{exp} and H_{exp} specify the range over which the atoms interact significantly. For example, it is well-known that the Si-Si interaction extends to only about 5Å. The remaining 8 parameters, 1 for each overlap and Hamiltonian integral, specify the strength or relative strength of the particular interaction. Although not as straightforward as the other parameters, initial values for these parameters can often be obtained from the existing literature (the relative strengths of the two-center interactions were widely used in early tight-binding calculations, which were often less computational and more analytical).

G Extended Hückel approximation I

The topic of this section, the Hückel approximation, appears in many different forms, some of which might not bear any overt resemblance to each other. Our discussion follows a non-standard development that is more suitable to the context of this report and to a modern computational treatment. If one considers the fact that both the overlap and Hamiltonian functions in eq. 1 represent interactions between atomic orbitals, one might consider the relation between these functions:

$$H_{ss\sigma}(R) = K_{ss\sigma}(R) \cdot S_{ss\sigma}(R)$$

This equation, which has a very similar form to the Hückel approximation, is at this point nothing more than a definition of the function $K_{ss\sigma}(R)$. Without making any approximations, one can set up the functions $K_{ss\sigma}(R)$, $K_{sp\sigma}(R)$, $K_{pp\sigma}(R)$, $K_{pp\pi}(R)$,

and then reformulate the model to consist of these 4 functions rather than the 4 Hamiltonian functions.

Next, if we consider the limiting value of $K_{ss\sigma}(R)$ for $R \rightarrow 0$ we have:

$$\lim_{R\rightarrow 0} H_{ss\sigma}(R) = \lim_{R\rightarrow 0} K_{ss\sigma}(R) \cdot \lim_{R\rightarrow 0} S_{ss\sigma}(R)$$

or:

$$\varepsilon'_s = \lim_{R\rightarrow 0} K_{ss\sigma}(R) \cdot 1$$

We can then reformulate the model again to consist of 4 unitless functions $k_{ss\sigma}(R)$ etc., each of which has a limiting value of 1 for $R \rightarrow 0$:

$$H_{ss\sigma}(R) = \varepsilon'_s \cdot k_{ss\sigma}(R) \cdot S_{ss\sigma}(R)$$

This reformulation becomes an approximation when one makes the argument that the 4 unitless functions $k_{ss\sigma}(R)$ etc. can be replaced with a single unitless function $k(R)$:

$$\begin{aligned} H_{ss\sigma}(R) &= \varepsilon'_s \cdot k(R) \cdot S_{ss\sigma}(R) \\ H_{sp\sigma}(R) &= \varepsilon'_{sp} \cdot k(R) \cdot S_{sp\sigma}(R) \\ H_{pp\sigma}(R) &= \varepsilon'_p \cdot k(R) \cdot S_{pp\sigma}(R) \\ H_{pp\pi}(R) &= \varepsilon'_p \cdot k(R) \cdot S_{pp\pi}(R) \end{aligned} \tag{11}$$

Note that in our development it is necessary to introduce a new parameter ε'_{sp} because the limiting values of $H_{sp\sigma}$ and $S_{sp\sigma}$ are both zero, leaving the limiting value of $K_{sp\sigma}$ undefined.

We have now arrived at one of the many variants of the extended Hückel approximation. The original development of this approximation is attributed to Anderson [13] and Hoffmann [14]; it was not until much later that the approximation was used in a model similar to our own by Menon [15]. In some early variants the function $k(R)$ is taken to be a constant $k(R) = 1$ for all R. In many variants the

prefactors ε' are constructed from the on-site energies ε using a single parameter K_0:

$$\varepsilon'_s \approx K_0 \cdot \varepsilon_s$$
$$\varepsilon'_{sp} \approx K_0 \cdot \tfrac{1}{2}(\varepsilon_s + \varepsilon_p)$$
$$\varepsilon'_p \approx K_0 \cdot \varepsilon_p$$
$$\varepsilon'_p \approx K_0 \cdot \varepsilon_p$$

This approximation was widely used in early less-computational calculations, with empirical values of K_0 from 1.75 to 2.25 [13].

If we consider these results in the context of our discussion of the limiting values of the Hamiltonian integrals in Section E, we have obtained a novel explanation of the Hückel approximation in terms of the behavior of the two-center integrals in eq. 1 for small values of R. This is a remarkable result, because all standard explanations of the Hückel approximation rely on a questionable argument that the atomic orbitals Φ are approximate eigenfunctions of the Hamiltonian operator $\hat{\mathrm{H}}$ with approximate eigenvalues ε (see Ref. [13]), which results in a questionable proportionality between the Hamiltonian and overlap matrix elements:

$$\int_{\mathbf{r}} \Phi_1(\mathbf{r}) \cdot \hat{\mathrm{H}} \cdot \Phi_2(\mathbf{r}) \cdot d\mathbf{r} \approx \varepsilon \cdot \int_{\mathbf{r}} \Phi_1(\mathbf{r}) \cdot \Phi_2(\mathbf{r})$$

In contrast, we have obtained such a proportionality based entirely on the limiting behavior of the two-center integrals for $R \to 0$. Most importantly, our explanation of the difference between the limiting values ε' and the on-site values ε in eq. 9 and 5 is consistent with the widely-used values of K_0 from 1.75 to 2.25. The widely-used extended Hückel approximation then, provides strong evidence for our argument of the importance of small values of R in empirical orbital modeling.

H Extended Hückel approximation II

In Section F we made the argument that there are enough similarities in the shapes of the overlap and Hamiltonian integrals that one can choose from a variety of constraints to reduce the number of empirical parameters. In this section we are

going to make a novel argument that the extended Hückel approximation can be interpreted as a particular set of constraints on the parameters of the overlap and Hamiltonian integrals. Let us begin by considering a highly simplified model where each of the 8 functions $S_{ss\sigma}(R)$ etc. has the functional form $e^{-\alpha \cdot R}$, i.e. a model with a total of 8 parameters $\alpha_{ss\sigma}$ etc.. Now, suppose that we are *required* to reduce the number of parameters from 8 to 1. The only reasonable choice would be to use the same value of α for each of the 8 functions. Next, suppose that we are required to reduce the number of parameters from 8 to 2. The important question now is whether we can find a *reasonable separation* of the 8 functions into 2 groups, so that we can use one parameter α for each group. There is of course such a reasonable separation; there is one group of overlap functions and one group of Hamiltonian functions. Note that this $8 \rightarrow 2$ case is very similar to one of our choices in Section F, where two *exp* parameters S_{exp} and H_{exp} are used for all 8 functions.

Continuing this line of reasoning, suppose that we are required to reduce the number of parameters from 8 to 4. Is there a reasonable separation of the 8 functions into 4 groups? There is indeed, it is just the separation into the groups $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$. With this particular separation, each overlap function is grouped with its corresponding Hamiltonian function; with this separation *we obtain a Hückel approximation* for this highly simplified model:

For illustrative purposes only. Do not attempt to use.

$$H_{ss\sigma}(R) = E_0 \cdot S_{ss\sigma}(R) = E_0 \cdot exp\,(-\alpha_{ss\sigma} \cdot R)$$
$$H_{sp\sigma}(R) = E_0 \cdot S_{sp\sigma}(R) = E_0 \cdot exp\,(-\alpha_{sp\sigma} \cdot R)$$
$$H_{pp\sigma}(R) = E_0 \cdot S_{pp\sigma}(R) = E_0 \cdot exp\,(-\alpha_{pp\sigma} \cdot R)$$
$$H_{pp\pi}(R) = E_0 \cdot S_{pp\pi}(R) = E_0 \cdot exp\,(-\alpha_{pp\pi} \cdot R)$$

Here we have included an energy prefactor E_0 solely for the purpose of having a Hamiltonian with the appropriate units. Using this line of reasoning, we can see that a Hückel approximation is one of many possible reasonable groupings of the overlap and Hamiltonian functions.

We have discussed groupings for $8 \rightarrow 1$, $8 \rightarrow 2$, $8 \rightarrow 4$, and $8 \rightarrow 8$ parameters. It might be useful to point out that one can use more complicated groupings to obtain almost any conceivable number of parameters. For example, one could argue that the $pp\sigma$ and $pp\pi$ functions should be grouped together because they both involve interactions between two p-orbitals. This leads to $8 \rightarrow 3$ and $8 \rightarrow 6$ groupings. As a final example, one could obtain a $8 \rightarrow 5$ Hückel approximation by using exponents α for the overlap functions and exponents $\alpha + \Delta\alpha$ for the Hamiltonian functions, i.e. with 4 different α's and only one $\Delta\alpha$. This corresponds to a widely used variant of the Hückel approximation where the function $k(R)$ in eq. 11 is taken to have the form $e^{-\Delta\alpha \cdot R}$.

It is clear that if we move away from this highly simplified model and return to the functional form in eq. 2, our line of reasoning, that the Hückel approximation is a particular set of parameter constraints, still holds. We could at this point attempt to specify exactly how the Hamiltonian parameters in eq. 2 can be reformulated to result in a proportionality between H and S that satisfies the Hückel approximation in eq. 11. For example, it is evident that if $H_{ss\sigma}(R)$ is to have an overall factor of ε'_s, one must reformulate not only $H_0 \rightarrow \varepsilon'_s \cdot H_0$ but also $H_1 \rightarrow \varepsilon'_s \cdot H_1$. However, such a reformulation is not useful or even necessary. The Hückel approximation relies on the usefulness of dividing a Hamiltonian function by its corresponding overlap function. This usefulness depends more on the shape of the functions, as in Figure 1, and less on a specific parameterization of the functions. In many ways we have already established this usefulness by our consideration of the shapes of the integrals in Figure 1 and the parameter constraints in Section F. The Hamiltonian and overlap functions have the same range of 5.0Å, the same crossover at 2.0Å, and the same relative strength. This is really all that is necessary to have a useful Hückel approximation.

In practice, if one uses a Hückel approximation, eq. 2 will not be used to construct the Hamiltonian functions, the Hamiltonian functions will be constructed

using eq. 11. We have used a Hückel approximation in much of our work on C, Si, and Ge. Our form for $k(R)$ follows that used by Menon in Ref. [15]:

$$k(R) = \exp\left(-K_{exp} \cdot R\right)$$

There does not appear to be any benefit to using a more complicated form for $k(R)$. This is likely due to the fact that the two-center integrals go to zero outsize a small range, and thus $k(R)$ is well-defined only for $R < 5.0$Å (using Si as an example). In our work K_{exp} is almost always small and negative during the fitting process, indicating that the Hamiltonian integrals have a slightly longer range than the overlap integrals. We should also point out that this form for $k(R)$ does not correspond exactly to using $H_{exp} = S_{exp} + K_{exp}$ for the *exp* parameter in eq. 2 because the *exp* parameter is involved other parts of eq. 2. In the context of the discussion in Section F, this leads us to another choice for a parameterization:

model with and $R \rightarrow 0$ and Hückel constraints:

---

3 parameters for each overlap function

0 parameters for each Hamiltonian function, plus ε'_s and ε'_p, plus ε'_{sp} and K_{exp}

on-site energies ε_s and ε_p

18 total parameters (two-center part only)

(12)

We have occasionally combined this parameterization with *exp* and *cross* constraints in eq. 10, but we usually return to the parameterization in eq. 12, which uses *exp* and *cross* parameters for each of the 4 overlap functions.

I Repulsive energy

With a parameterized form for the 8 functions for the overlap and Hamiltonian in eq. 1, it remains to specify a parameterized form for the repulsive energy function $E_{rep}(R)$, which is used to construct the repulsive energy E_{rep} for a

system of atoms indexed by i and j as:

$$E_{rep} = \sum_{i,j} E_{rep}(R_{ij})$$

This simple two-body or pairwise energy is added to the band energy after the eigenvalue equation has been solved, i.e. this energy is not involved in the eigenvalue equation. In the formalism of first-principles or ab-initio approaches, the total energy E_{tot} consists of three combinations of interactions between nuclei and electrons:

$$E_{tot} = E_{electrons\text{-}nuclei} + E_{nuclei\text{-}nuclei} + E_{electrons\text{-}electrons}$$

The band energy E_{band} accounts for the interaction between electrons and nuclei, but for mathematical reasons E_{band} also contains an unavoidable "double-counting" of the energy between electrons and other electrons:

$$E_{band} = E_{electrons\text{-}nuclei} + 2 \cdot E_{electrons\text{-}electrons}$$

This results in an expression for the total energy E_{tot} in which the energy $E_{e\text{-}e}$ appears with an explicit negative sign:

$$E_{tot} = E_{band} + E_{nuclei\text{-}nuclei} - E_{electrons\text{-}electrons} \tag{13}$$

The repulsive energy, or more appropriately the energy not accounted for by the band energy, is then:

$$E_{non\text{-}band} = E_{nuclei\text{-}nuclei} - E_{electrons\text{-}electrons}$$

This result follows very closely a discussion by Chadi [4].

The traditional argument for replacing the very complicated first-principles energy $E_{non\text{-}band}$ with the very simple empirical energy E_{rep} is that the energies $E_{n\text{-}n}$ and $E_{e\text{-}e}$, which are both long-range, under certain conditions combine to give a short-range energy that is repulsive and can be constructed from a pairwise energy $E_{rep}(R)$. However, although one can say that the resulting empirical model is

accurate, it is apparently very dangerous to make claims about the accuracy of the intermediate steps used to arrive at E_{rep}. In a detailed analysis on this topic, Foulkes and Haydock [16] compared tight-binding (TB) to density-functional theory (DFT) and concluded:

> The origin of [the TB expression for the total energy] is not at all clear. It looks rather like [the DFT expression for the total energy], but the double-counting (and nuclear-nuclear repulsion) terms are now assumed to be pairwise and short-ranged (which is certainly not the case if charge transfer leads to long-range interatomic Coulomb forces) and the [energy eigenvalues] are now the solutions of a non–self-consistent Schrödinger equation rather than a self-consistent one. It seems, therefore, that [the TB expression for the total energy] ignores self-consistency and assumes that all the important nonpairwise behavior in the interatomic forces comes from the sum of the one-electron eigenvalues. In fact, as we will explain, neither of these conclusions is quite right and the approximations behind [the DFT expression for the total energy] are rather more subtle and sophisticated than they appear. [16]

There have been several attempts to develop empirical orbital models with more elaborate repulsive energies, i.e. more elaborate than a two-center or pairwise function $E_{rep}(R)$. For example, Mercer and Chou [8] include higher-order energy terms that depend on the angles associated with three atoms. We prefer to think of such higher-order terms as modifications to models which consist strictly of the 9 functions in eq. 1. There are two reasons for this. First, the broader context of our report is the development of *environment-dependent* models. In this context, these more elaborate repulsive energies look very much like specific cases of environment-dependence. Our environment-dependent model does not even *have* a repulsive energy; it treats the total energy using an expression similar to eq. 13.

The second reason is that there does not appear to be any "standard extension", either in theory or in practice, to a pairwise repulsive energy. Various extensions using bond angles, coordination numbers, and atomic charges are all in current use [8]. In an empirical orbital model, the only standard form for the repulsive energy is a two-center pairwise form.

When we need to use a self-contained two-center model, i.e. a model with no environment-dependent modifications, we use the following simple form for the repulsive energy:

$$E_{rep}(R) = (E_0 + E_1 \cdot R) \cdot \exp(-E_{exp} \cdot R)$$

The parameter E_1 should be relatively small, or more appropriately, should have a relatively small effect on the shape of the function; the repulsive energy is always positive, it does not oscillate. The range of $E_{rep}(R)$ is apparently always significantly smaller than the range of the overlap and Hamiltonian integrals, i.e. the exponent E_{exp} should be significantly larger than S_{exp} and H_{exp}. We have observed this behavior across a large range of empirical parameter fittings for C, Si, and Ge. This is the same behavior that is observed by Foulkes and Haydock [16] and by Frauenheim [9], [10], [11] by extracting a pairwise energy from density-functional theory calculations.

Unlike the overlap and Hamiltonian functions, the repulsive energy does not appear to contain any useful information for any values of R smaller than those which might be observed experimentally. There is no benefit to using a functional form such as $\frac{1}{R} \cdot e^{-E_{exp} \cdot R}$ that has a more appropriate behavior for $R \to 0$. The repulsive energy should be interpreted as being defined only for experimentally relevant values of R, or for values of R used during the fitting process (as it is often desirable to use smaller-than-experimental values for fitting). We have also observed a behavior that has not been reported in the literature. The calculated values of the fitting properties do not depend strongly on any properties of the repulsive energy other than the value of the repulsive energy at the nearest-neighbor distance of the

material. This suggests that the repulsive energy might do little more than count the number of nearest-neighbors:

$$E_{tot} = E_{band} + N_{coord} \cdot E_{coord}$$

where N_{coord} is some average coordination number and E_{coord} is some average coordination energy.

CHAPTER III

NEXT-GENERATION MODELS

A Too many functions

Our parameterization of the 9 functions in eq. 1 ends with our discussion of the repulsive energy in the previous section. In this chapter I will discuss concepts for what might be called the "next generation" of two-center techniques. My interest in these concepts grew out of a frustration with a particular mathematical "feature" of the overlap matrix that results in a failure, both in theory and in practice, to solve the eigenvalue equation for a system of atoms. I will discuss this issue of non-positive-definite overlap in detail in this chapter. However, the resulting concepts are best tied together by the fact that they address the issue of applying two-center techniques to more complicated materials, and it is with this issue that we will begin.

It is becoming clear from the recent literature, conferences, and also from recent trends in funding, that several features will be demanded of the next generation of empirical orbital models. First, the model *must treat multi-element systems natively.* Although there will still be important applications that involve only a single type of atom, such as carbon nanotubes and silicon surfaces, such applications have become marginalized by the demand for problems with more than one type of atom. Similarly, the model must be able to treat the most important elements in biochemistry and pharmaceuticals. Since most models are already able to treat hydrogen and carbon, what this really means is that the model *must be able to treat nitrogen and oxygen.* Similarly, the model *must treat d-orbitals and*

near-valence orbitals natively. These orbitals are needed for the transition metals, the calculation of optical and spectroscopic properties, and (arguably) to improve the accuracy of s, p orbital calculations.

Next, the model *must treat the calculation of properties other than the energy natively.* Traditionally, the core of material science calculations has been the energy landscape, i.e. the energy as a function of the coordinates of the nuclei. From this one can calculate equilibrium energies and geometries, forces and elastic coefficients, band structures (if individual energy eigenvalues are included), and also a very large variety of kinetic and thermal properties. However, this core is arguably being replaced by the ever-increasing need to calculate properties that can not be obtained from a knowledge of the energy landscape only. The electron density $\rho(\mathbf{r})$ is a representative example of such a property. Finally, I will add to this list my own requirement that *all reasonable configurations of atoms must have a calculatable energy.* This requirement is related to the issue of non-positive-definite overlap, which will be discussed later.

At this point in our discussion, it is not clear that models based on the 9 functions in eq. 1 do not already satisfy these requirements. In fact, empirical orbital models are widely used for multi-element systems, d-orbitals and near-valence orbitals, nitrogen and oxygen, and not-just-energy properties. I will make the argument in this report that models based on the 9 functions in eq. 1 do not satisfy these requirements. To begin this argument, let us consider a system consisting of the elements H, C, N, O, and Fe. This is intended to represent a biological system; iron has been chosen to illustrate the effect of d-orbitals. The standard or minimal basis set that one would use consists of s orbitals for hydrogen; s, p orbitals for carbon, nitrogen, and oxygen; and s, p, d orbitals for iron. How many functions, corresponding to the 9 functions in eq. 1, must be specified for this system? For the H-H interaction there is only one function $ss\sigma$. For the C-C, N-N, and O-O interactions there are $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$. For the Fe-Fe interaction there

H-H	$ss\sigma$		
C-C	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		
N-N	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		
O-O	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		
Fe-Fe	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$	$sd\sigma$, $pd\sigma$, $pd\pi$, $dd\sigma$, $dd\pi$, $dd\delta$	
H-C	$ss\sigma$, $sp\sigma$		
H-N	$ss\sigma$, $sp\sigma$		
H-O	$ss\sigma$, $sp\sigma$		
C-N	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		$ps\sigma$
C-O	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		$ps\sigma$
N-O	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$		$ps\sigma$
H-Fe	$ss\sigma$, $sp\sigma$, $sd\sigma$		
C-Fe	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$	$sd\sigma$, $pd\sigma$, $pd\pi$	$ps\sigma$
N-Fe	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$	$sd\sigma$, $pd\sigma$, $pd\pi$	$ps\sigma$
O-Fe	$ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$	$sd\sigma$, $pd\sigma$, $pd\pi$	$ps\sigma$

Figure 3. Tabulation of the functions that must be specified for a system consisting of the five elements H, C, N, O, Fe.

are ten functions $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$, plus $sd\sigma$, $pd\sigma$, $pd\pi$, $dd\sigma$, $dd\pi$, $dd\delta$.

Next, we must consider the H-C, H-N, and H-O interactions. Each consists of only two functions $ss\sigma$ and $sp\sigma$ ($pp\sigma$ and $pp\pi$ are not present here because p orbitals are not used for hydrogen). Next, the C-N, C-O, and N-O interactions each consist of the four functions $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$. However, it turns out that one must also include a fifth function $ps\sigma$. Such additional functions are necessary in general for interactions between two different elements. For example, the $sp\sigma$ function for C-N represents an interaction between two different types of atoms (carbon and nitrogen) and two different types of orbitals (s and p), and thus requires a corresponding function $ps\sigma$. Turning now to the Fe interactions, for H-Fe we have only three functions $ss\sigma$, $sp\sigma$, $sd\sigma$. Again, functions such as $pd\sigma$ and $dd\sigma$ are not present for hydrogen. Finally, for C-Fe, N-Fe, and O-Fe we have all the functions that consist of s, p in the first position and s, p, d in the second position. This includes $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$, plus $sd\sigma$, $pd\sigma$, $pd\pi$, plus the corresponding function $ps\sigma$. These results are tabulated in Figure 3.

It is evident from Figure 3 that we have a problem. There are *too many*

functions. If we account for both the overlap and Hamiltonian interactions, there are 142 functions for a system consisting of only five different types of atoms. To be fair, we should point out that in practice many of these interactions can be taken to be zero, either because a particular element is known not to bond to another particular element, or because the orbitals involved in the interaction are known to interact weakly. We should also emphasize that it is never intended to fit all these functions simultaneously. In the best-case scenario, each pair of elements would be fit separately, i.e. one fitting for each row in Figure 3, and so the dimensionality of the optimization problem is affected only by the number of functions in each row, and not by the total number of functions in all rows. However, starting with a s,p model for a single element, it is clear from this table that adding a different element, or adding d-orbitals or near-valence orbitals, results in what might be called an explosion in the number of functions and empirical parameters. The interaction between two different transition metals requires a whopping *fourteen* functions $ss\sigma$, $sp\sigma$, $pp\sigma$, $pp\pi$, plus $sd\sigma$, $pd\sigma$, $pd\pi$, $dd\sigma$, $dd\pi$, $dd\delta$, plus $ps\sigma$, $ds\sigma$, $dp\sigma$, $dp\pi$, not including the ten functions each for the interactions between the same type of atom.

From this we can make the argument that a model based on the 9 functions in eq. 1 was never designed for such systems, similarly that if asked to develop a model for such systems from scratch, one would not develop the current model, and similarly that the model does not treat these systems *natively.* Before proceeding, it is useful to briefly discuss two issues that help put this problem in context. The first issue is that of *averaging.* There is a general feeling in the community, which some might even call an axiom, that a model for multi-element systems should be able to be constructed from the individual models for each single-element system, i.e. with little or no additional parameter fitting. This would mean that any multi-element interaction could be constructed from the corresponding single-element interactions, using some type of averaging scheme, for example:

$$\text{C-N}_{ss\sigma} = \tfrac{1}{2} \cdot (\text{C-C}_{ss\sigma} + \text{N-N}_{ss\sigma}) \tag{14}$$

In this context, we can reformulate the requirement for multi-element systems to say that the model *must treat multi-element averaging natively.* Models based on the 9 functions in eq. 1 definitely do not treat averaging natively.

The second issue is that of *first-principles-style models.* This is related to the previously-mentioned technique of discarding functions based on a prior knowledge that certain interactions either do not occur or are weak. Although often very useful, such techniques work against the concept of having as arbitrary or as general a model as possible. This arbitrariness is a very well-liked feature of first-principles models; one can ask for a calculation on almost any configuration of atoms no matter how exotic. This feature is so well-liked that it is becoming expected of empirical orbital models. For example, if one discards the Fe-Fe interactions, one can treat systems where Fe atoms are known not to bond to other Fe atoms. However, such a model could never be used to study surfaces of crystalline iron.

This issue of arbitrariness is closely related to the historical development of empirical orbital models. Following our previous discussion, it was not until the mid 1990s that the overlap and Hamiltonian functions came to be regarded as arbitrary or general functions of a scalar variable R. In earlier calculations, only the values at the first few nearest-neighbor distances were used. These earlier models actually treated multi-element systems more naturally or more natively than the later models. This is because the identification of nearest-neighbor values, combined with other techniques such as hybridization and crystal symmetry, did not result in an explosion in the number of functions and parameters. The result is something of a paradox, in that the earlier models are in some ways more adept at complex systems than the later models. The important point is that the treatment of multi-element systems becomes more problematic, not less, as one takes the more modern approach of treating the overlap and Hamiltonian functions as arbitrary functions of R.

B Non-positive-definite overlap

The first sign of trouble was that our source code was reporting errors (or "exceptions" in the language of software engineering), specifically that for certain sets of empirical parameters the fitting properties could not be evaluated. The program's error handling and error reporting features (essential features for any large program) traced the problem to the failure of the eigenvalue equation solver to calculate the energy eigenvalues. The eigenvalue equation solver was reporting that the overlap matrix was not positive definite. At first this did not seem to be a problem, as it simply meant that the offending sets of parameters needed to be discarded (which they were). However, I became convinced over time that this was indeed a serious problem. Although this exception often occurred for the more "extreme" fitting properties, such as those with unusually small bond lengths, it also occurred for some likely experimentally observable properties. Also, we would find that a best or optimized set of parameters, i.e. a set that worked during the fitting process, would sometimes not work when applied to other configurations of atoms. We have also observed this behavior for other models similar to our own, where the parameters for these models are available in the published literature.

The problem of non-positive-definite overlap is not just a low-level computational problem. It is a high-level theoretical problem. The eigenvalue equation that results from using a non-orthogonal basis set is:

$$H_{i\alpha,j\beta} \cdot C_{j\beta,\lambda} = S_{i\alpha,j\beta} \cdot C_{j\beta,\lambda} \cdot E_{\lambda}$$

or in matrix form:

$$\mathrm{H} \cdot \mathrm{C} = \mathrm{S} \cdot \mathrm{C} \cdot \mathrm{E}$$

where H and S are the Hamiltonian and overlap matrix, C is the eigenvector matrix, and E is the eigenvalue array. Now, for the *orthogonal* problem, it is well-known that if H is Hermitian the energies E are guaranteed to be real. This is so important that it is considered to be axiomatic that any modeled Hamiltonian must be Hermitian.

However, for the non-orthogonal problem, if H is Hermitian then the energies E are not guaranteed to be real. This can be seen in the context of the orthogonal problem by constructing the orthogonalized Hamiltonian $V = S^{-1} \cdot H$ which results in the orthogonal equation $V \cdot C = C \cdot E$. Even with both H and S Hermitian, V is still not Hermitian, and the energies E can not be guaranteed to be real.

In the context of the non-orthogonal problem, one avoids taking the inverse of S and instead constructs the *Cholesky factorization* U of the overlap:

$$S = U^{\dagger} \cdot U$$

where U is an upper triangular matrix. The non-orthogonal problem can then be cast in the orthogonal form [17]:

$$\left((U^{-1})^{\dagger} \cdot H \cdot U^{-1}\right) \cdot (U \cdot C) = (U \cdot C) \cdot (E)$$

The cast Hamiltonian $(U^{-1})^{\dagger} \cdot H \cdot U^{-1}$ is guaranteed to be Hermitian, and the energies E are guaranteed to be real. However, the crux of the matter is that the Cholesky factorization U exists only if the overlap matrix S is positive definite. In fact, the existence of U can be taken to define whether a (symmetric) matrix is positive definite. The important result is that for the non-orthogonal problem, one can have a Hermitian Hamiltonian and still have configurations of atoms that *do not have a calculatable energy.* Our calculatable energy requirement in the previous section can then be reformulated to say that *the overlap matrix must be positive definite.* Undesirable non-positive definite overlap is a fundamental feature of models which use the 9 functions in eq. 1, as we will discuss in the next section.

The lack of requiring the overlap matrix to be positive definite is arguably a flaw or oversight in the historical development of empirical orbital models. Although tight-binding models date back to Slater's 1954 paper, the importance of using a non-orthogonal Hamiltonian was not recognized until a 1993 paper by Canel, Carlsson, and Fedders [18], and it was not until the late 1990s that non-orthogonal models were relatively widely used. In all models it was always taken as axiomatic

that the Hamiltonian matrix was required to be Hermitian. However, the main reason for this requirement is to guarantee that the energy eigenvalues are real. As a transition was made to non-orthogonal models, the Hamiltonian was still required to be Hermitian, but this requirement is somewhat pointless if the overlap matrix is not also required to be positive definite. It is interesting to note that orthogonal models, which have a longer history and have been more widely used, already satisfy our requirement of positive definite overlap. This is of course because for orthogonal models, the overlap matrix is just the identity matrix, which is always positive definite.

C Integrability constraints I

How is it possible for the 9 functions in eq. 1 to result in systems with no calculatable energy? More specifically, what are the theoretical properties of overlap matrices that are responsible for maintaining positive definite overlap? In this section we will show that it is the construction of overlap matrix elements as actual *integrals* of actual *atomic orbitals* that maintains positive definite overlap, and that the loss of this property is responsible for systems which do not have a calculatable energy. It is useful to begin with an informal argument based on the "degrees of freedom" involved in the overlap matrix. For the empirical model the 4 functions $S_{ss\sigma}(R)$, $S_{sp\sigma}(R)$, $S_{pp\sigma}(R)$, $S_{pp\pi}(R)$ can be interpreted as 4 degrees of freedom; in the most general case these functions are parameterized independently of each other. Each degree of freedom is a scalar function of a scalar variable R defined for values of R from 0 to ∞. However, the functions $S_{ss\sigma}(R)$ etc. are intended to represent integrals of atomic orbitals:

$$\begin{aligned} S_{ss\sigma}(R) &= \int_{\mathbf{r}} \Phi_s(\mathbf{r}-\mathbf{R}_1)\cdot\Phi_s(\mathbf{r}-\mathbf{R}_2)\cdot d\mathbf{r} \\ S_{sp\sigma}(R) &= \int_{\mathbf{r}} \Phi_s(\mathbf{r}-\mathbf{R}_1)\cdot\Phi_p(\mathbf{r}-\mathbf{R}_2)\cdot d\mathbf{r} \\ S_{pp\sigma}(R) &= \int_{\mathbf{r}} \Phi_p(\mathbf{r}-\mathbf{R}_1)\cdot\Phi_p(\mathbf{r}-\mathbf{R}_2)\cdot d\mathbf{r} \\ S_{pp\pi}(R) &= \int_{\mathbf{r}} \Phi_p(\mathbf{r}-\mathbf{R}_1)\cdot\Phi_p(\mathbf{r}-\mathbf{R}_2)\cdot d\mathbf{r} \end{aligned} \qquad \text{with } R = \|\mathbf{R}_2-\mathbf{R}_1\| \qquad (15)$$

where the symmetry notation σ and π refers to the relative orientation of the orbitals at positions $\mathbf{R}_1$ and $\mathbf{R}_2$: for the p-orbitals, σ refers to a p_z orbital oriented along the same axis as $\mathbf{R}_2 - \mathbf{R}_1$, while π refers to a p_x or p_y orbital oriented along the same axis as $\mathbf{R}_2 - \mathbf{R}_1$(of course for the s-orbitals there is only one possible orientation, σ). How many degrees of freedom are there if these integrations are performed explicitly? There are 4 atomic orbitals $\Phi_s(\mathbf{r})$, $\Phi_{p_x}(\mathbf{r})$, $\Phi_{p_y}(\mathbf{r})$, $\Phi_{p_z}(\mathbf{r})$. The angular parts of these three-dimensional functions are fixed, and there are only 2 independent radial functions, or degrees of freedom, $\phi_s(r)$ and $\phi_p(r)$, which are defined for values of r from 0 to ∞. Without the integration then, there are 4 degrees of freedom, but with the integration there are only 2 degrees of freedom.

This conflict in the number of degrees of freedom is more dramatic if one considers the d-orbitals. If one adds d-orbitals to an existing s and p orbital model, there is only one new radial function $\phi_d(r)$. However, there are *six* new two-center integrals $S_{sd\sigma}(R)$, $S_{pd\sigma}(R)$, $S_{pd\pi}(R)$, $S_{dd\sigma}(R)$, $S_{dd\pi}(R)$, $S_{dd\delta}(R)$. This is an example of the explosion in the number of functions discussed in Section A. The conflict is even more dramatic if one considers the five-element example in Figure 3. Here, there is 1 radial function for hydrogen, 2 each for carbon, nitrogen, and oxygen, and 3 for iron, for a total of 10 radial functions or degrees of freedom. There are a whopping *seventy-one* degrees of freedom Figure 3.

This can be stated more formally by saying that the integrals in eq. 15 are *convolutions* of atomic orbitals, and that the atomic orbitals are *deconvolutions* of the integrals. That is, the convolutions are mathematical operations that map input functions $\phi(r)$ to output functions $S(R)$, and the deconvolutions map input functions $S(R)$ to output functions $\phi(r)$. It is these deconvolutions that show that there are indeed theoretical conflicts in treating the $S(R)$ as independent functions.

First, let us express the integrals in eq. 15 in functional notation:

$$S_{ss\sigma}(R) = S\left[\phi_s(r)\right]$$
$$S_{sp\sigma}(R) = S\left[\phi_s(r), \phi_p(r)\right]$$
$$S_{pp\sigma}(R) = S\left[\phi_p(r)\right]$$
$$S_{pp\pi}(R) = S\left[\phi_p(r)\right]$$

where the brackets indicate a functional dependence. If we consider the expression for $S_{pp\sigma}(R)$, this implies that the radial function $\phi_p(r)$ can be constructed entirely from a knowledge of the two-center integral $S_{pp\sigma}(R)$:

$$\phi_p(r) \overset{\text{deconv.}}{=} \phi\left[S_{pp\sigma}(R)\right] \tag{16}$$

This deconvolution shows that the existence of a well-defined two-center integral $S_{pp\sigma}(R)$ implies the existence of a well-defined radial function $\phi_p(r)$. However, the expression for $pp\pi$ implies the existence of a *different* radial function:

$$\phi_p(r) \overset{\text{deconv.}}{=} \phi\left[S_{pp\pi}(R)\right] \tag{17}$$

The radial functions in eq. 16 and 17 will be different if the two-center integrals $S(R)$ are treated as independent functions. The tight-binding model, i.e. the 9 functions in eq. 1, *implies the existence of multi-valued radial functions.* This is also the case for $\phi_s(r)$. For $ss\sigma$ we have the deconvolution:

$$\phi_s(r) \overset{\text{deconv.}}{=} \phi\left[S_{ss\sigma}(R)\right] \tag{18}$$

For $sp\sigma$ the situation is slightly more complicated. First, there is a partial or mixed deconvolution:

$$\phi_s(r) \overset{\text{deconv.}}{=} \phi\left[S_{sp\sigma}(R), \phi_p(r)\right]$$

But we can substitute for $\phi_p(r)$ from eq. 16 to obtain:

$$\phi_s(r) \overset{\text{deconv.}}{=} \phi\left[S_{sp\sigma}(R), S_{pp\sigma}(R)\right] \tag{19}$$

The radial functions $\phi_s(r)$ in eq. 18 and 19 are in general different. These results clearly extend to other orbitals, such as d-orbitals and excited or near-valence orbitals.

It is actually a straightforward matter to show that the existence of well-defined or single-valued radial functions guarantees that the overlap matrix is always positive definite, and hence guarantees that a system has a calculatable energy. Start with the definition [17] of a positive definite matrix S:

$$\mathrm{x} \cdot \mathrm{S} \cdot \mathrm{x} > 0$$

where x is any array. Express the overlap matrix S in indexed form:

$$S_{i\alpha,j\beta} = \int_{\mathbf{r}} \phi_\alpha(\mathbf{r} - \mathbf{R}_i) \cdot \phi_\beta(\mathbf{r} - \mathbf{R}_j) \cdot d\mathbf{r}$$

where i and j index the atomic nuclei, and α and β index the atomic orbitals at each nucleus. Note that although i and α are separate indexes for the purposes of the integration, $i\alpha$ is a single index for the purposes of the eigenvalue equation. This means that the array $\mathbf{x}$ is indexed by $i\alpha$. This gives:

$$\sum_{i\alpha,j\beta} x_{i\alpha} \cdot x_{j\beta} \cdot S_{i\alpha,j\beta} = \int_{\mathbf{r}} \left(\sum_{i\alpha} x_{i\alpha} \cdot \phi_\alpha(\mathbf{r} - \mathbf{R}_i)\right) \cdot \left(\sum_{j\beta} x_{j\beta} \cdot \phi_\beta(\mathbf{r} - \mathbf{R}_j)\right) \cdot d\mathbf{r}$$

Whatever the item $\sum_{i\alpha} \ldots \cdot \sum_{j\beta} \ldots$ is, it is something squared, which is always positive. This important result proves our earlier claim that it is the actual integration of actual atomic orbitals that maintains positive definite overlap and calculatable energies.

D Integrability constraints II

The requirement of positive definite overlap suggests that we move away from an "overlap-parameterized" model, and toward a model which features a direct parameterization of the radial functions of the atomic orbitals. Our research in this area took place towards the end of our project, and so for logistical reasons we have

not significantly tested such a model. Without this testing, we will refer to direct parameterization of the radial functions as a *prototype*, and our discussion will focus on broader issues rather than on specific parameterizations of the radial functions. We can see from our discussion in the previous section that this prototype satisfies the requirement of positive definite overlap, and thus also satisfies the requirement that all reasonable configurations of atoms have a calculatable energy.

Next, we can see that this prototype satisfies the requirement that multi-element systems are treated natively, and that d-orbitals and near-valence orbitals are also treated natively. This is due to the fact that treating the radial functions as degrees of freedom does not result in an explosion in the number of empirical parameters as one moves from a single-element system to a multi-element system, or as one adds d-orbitals or near-valence orbitals to an existing s, p orbital model. As we have already discussed, in Figure 3 there are seventy-one independent functions for the overlap-parameterized model, but only 10 independent functions for the radial-function prototype.

The key to understanding how this is effected is to consider multi-element averaging (see eq. 14). For example, if one has a radial-function parameterization for C-C, and a separate radial-function parameterization for N-N, this means that one has the radial functions $\phi_s^{\mathrm{C}}(r)$ for carbon and $\phi_s^{\mathrm{N}}(r)$ for nitrogen. Then, for the C-N interaction, we will have

$$S_{ss\sigma}^{\mathrm{C-N}}(R) = \int_{\mathbf{r}} \phi_s^{\mathrm{C}}(\mathbf{r} - \mathbf{R}_i) \cdot \phi_s^{\mathrm{N}}(\mathbf{r} - \mathbf{R}_j) \cdot d\mathbf{r} \tag{20}$$

With an overlap-parameterized model, $S_{ss\sigma}^{\mathrm{C-N}}$ would either need to be parameterized separately, or would need to be constructed from $S_{ss\sigma}^{\mathrm{C}}$ and $S_{ss\sigma}^{\mathrm{N}}$ using a questionable averaging scheme as in eq. 14. With the radial-function prototype, no additional parameterization is necessary, and there is no questionable averaging. If one wants to interpret eq. 20 as a type of averaging, then we have the result that the radial-function prototype satisfies the requirement that multi-element averaging is treated natively.

We have four important items to mention about such a prototype. The first item concerns the construction of the two-center integrals $S_{ss\sigma}(R)$ etc.. We should clarify that, for a radial-function model, these functions will still be constructed. This is due to the fact that the functions $S_{ss\sigma}(R)$ etc., regardless of whether they are parameterized directly or not, provide the fastest way to calculate the elements of the overlap and Hamiltonian matrix. In fact, Slater's 1954 paper, which established the parameterization of these functions, also established a recipe for using these functions to construct the overlap and Hamiltonian matrix. For example:

$$S_{x,x} = l^2 \cdot S_{pp\sigma} - (1 - l^2) \cdot S_{pp\pi} \tag{21}$$

specifies how to constrict the matrix element $S_{i\alpha,j\beta}$ when α and β refer to p_x orbitals. Further discussion of these formulas is not needed, other than to point out that this recipe still appears to be the best way to construct the matrix elements for our prototype.

The second issue is that in some cases it is possible to perform the integration over the radial functions analytically, making it possible to provide explicit functional forms for $S_{ss\sigma}(R)$ etc.. In Figure 4 we show analytical forms for $S_{ss\sigma}(R)$, $S_{sp\sigma}(R)$, $S_{pp\sigma}(R)$, $S_{pp\pi}(R)$ obtained using the parameterized forms for the radial functions:

$$\begin{aligned} \phi_s(r) &= (S_0 + S_1 \cdot r + S_2 \cdot r^2) \cdot exp(-E \cdot r) \\ \phi_p(r) &= (P_0 + P_1 \cdot r + P_2 \cdot r^2) \cdot exp(-E \cdot r) \end{aligned} \tag{22}$$

Although these integrations are so unwieldy that one would never attempt to do them on paper, one can use a computer algebra software package, which we did to obtain the results in Figure 4 (this is not a straightforward matter, as the multi-dimensional integrals require a transformation to prolate two-center spheroidal coordinates). This leads to a remarkable conclusion that our prototype is in some ways nothing more than a complicated parameterization of the 9 functions in eq. 1, with a complicated set of parameter constraints as in Section F. That is, a

very special set of constraints that guarantees positive definite overlap and provides for native multi-element averaging. Even in the general case where analytical integration is not possible, this interpretation is still meaningful in that one still has a parameterization (although not analytical) of the 9 functions in eq. 1.

The third issue is the behavior of the two-center integrals for small values of R. In Chapter II we discussed in detail the importance of treating the two-center integrals $S_{ss\sigma}(R)$ etc. for all values of R, including small values. Our interest in integrability constraints developed largely independently of our interest in small values of R. However, it turns out that the two topics are related in an important way. The concept that the existence of two-center integrals implies the existence of radial functions, as in eq. 16, formally requires that the two-center integrals $S_{ss\sigma}(R)$ etc. are defined for all values of R. This is due to the fact that the functional dependence indicated by the brackets in eq. 16 means that the value of $\phi_p(r)$ at just one specific value of r depends on the values of $S_{pp\sigma}(R)$ at all the values of R. I would like to emphasize that, as an *empirical* model, one can make a strong argument for a radial function model without this formal requirement. It is still interesting to note that a radial function model is consistent with a treatment of small values of R.

The fourth item concerns the global optimization problem. We have discussed previously that the search for the best set of parameters involves a poorly behaved optimization function, with a large number of poorly distributed local minima. Now, in general such poor behavior *might* be an unfortunate but inherent part of an optimization problem. However, such behavior can also indicate that an optimization problem has too many degrees of freedom, i.e. that there are hidden or unaccounted-for constraints. There is indeed evidence that the poor behavior of our optimization function is *not* an inherent part of the problem: we have observed that many local minima have very similar patterns as to how the calculated values differ from the reference values. For example, if one local minimum gives a bond length

$$S_{ss\sigma}(R) = \tfrac{1}{840} \cdot \tfrac{1}{E^7} \cdot e^{-E\cdot R} \cdot ($$
$$3E^6S_2^2R^6 + 21E^5S_2^2R^5 + 14S_2E^6S_1R^5 + 14E^6S_1^2R^4 + 84E^5S_1S_2R^4$$
$$+126E^4S_2^2R^4 + 42E^6S_2S_0R^4 + 210E^5S_0S_2R^3 + 70E^5S_1^2R^3 + 70E^6S_1S_0R^3$$
$$+420E^4S_1S_2R^3 + 630E^3S_2^2R^3 + 1470E^3S_1S_2R^2 + 2205E^2S_2^2R^2$$
$$+630E^4S_0S_2R^2 + 70E^6S_0^2R^2 + 280E^4S_1^2R^2 + 280E^5S_1S_0R^2$$
$$+1260E^3S_0S_2R + 4725ES_2^2R + 3150E^2S_1S_2R + 210E^5S_0^2R$$
$$+630E^4S_0S_1R + 630E^3S_1^2R + 1260E^2S_0S_2 + 630E^2S_1^2 + 630S_0S_1E^3$$
$$+3150S_1S_2E + 4725S_2^2 + 210S_0^2E^4\ )$$

$$S_{sp\sigma}(R) = -\tfrac{\sqrt{3}}{840} \cdot \tfrac{1}{E^6} \cdot R \cdot e^{-E\cdot R} \cdot ($$
$$3E^5P_2S_2R^5 + 17E^4P_2S_2R^4 + 7E^5P_1S_2R^4 + 7E^5P_2S_1R^4 + 77E^3P_2S_2R^3$$
$$+21E^5P_2S_0R^3 + 35E^4P_2S_1R^3 + 14E^5P_1S_1R^3 + 21E^5P_0S_2R^3 + 28E^4P_1S_2R^3$$
$$+91E^4P_2S_0R^2 + 35E^5P_1S_0R^2 + 49E^4P_1S_1R^2 + 252E^2P_2S_2R^2 + 63E^3P_1S_2R^2$$
$$+35E^5P_0S_1R^2 + 140E^3P_2S_1R^2 + 21E^4P_0S_2R^2 + 525ES_2P_2R + 35E^4P_0S_1R$$
$$+105E^3P_1S_1R + 315E^2P_2S_1R + 105E^4P_1S_0R + 70E^5S_0P_0R + 210E^3P_2S_0R$$
$$+105E^2P_1S_2R + 525S_2P_2 + 105E^2P_1S_1 + 105EP_1S_2 + 35E^3P_0S_1$$
$$+210E^2P_2S_0 + 105E^3P_1S_0 + 70E^4S_0P_0 + 315EP_2S_1\ )$$

$$S_{pp\sigma}(R) = -\tfrac{1}{280} \cdot \tfrac{1}{E^7} \cdot e^{-E\cdot R} \cdot ($$
$$3E^6P_2^2R^6 + 14E^6P_1P_2R^5 + 13E^5P_2^2R^5 + 34E^4P_2^2R^4 + 42E^5P_1P_2R^4$$
$$+14E^6P_1^2R^4 + 42E^6P_2P_0R^4 + 14E^4P_1P_2R^3 + 28E^5P_1^2R^3 + 14E^5P_0P_2R^3$$
$$-42E^3P_2^2R^3 + 70E^6P_1P_0R^3 - 567E^2P_2^2R^2 - 336E^3P_1P_2R^2 + 70E^6P_0^2R^2$$
$$-126E^4P_2P_0R^2 - 42E^4P_1^2R^2 - 1575EP_2^2R - 420E^3P_0P_2R - 210E^3P_1^2R$$
$$-1050E^2P_1P_2R - 210E^4P_0P_1R - 70E^5P_0^2R - 1050P_1P_2E - 1575P_2^2$$
$$-420E^2P_0P_2 - 210E^2P_1^2 - 70P_0^2E^4 - 210P_0P_1E^3\ )$$

$$S_{pp\pi}(R) = \tfrac{1}{280} \cdot \tfrac{1}{E^7} \cdot e^{-E\cdot R} \cdot ($$
$$3E^5P_2^2R^5 + 14E^5P_1P_2R^4 + 31E^4P_2^2R^4 + 112E^4P_1P_2R^3 + 14E^5P_1^2R^3$$
$$+42E^5P_0P_2R^3 + 189E^3P_2^2R^3 + 462E^3P_1P_2R^2 + 84E^4P_1^2R^2 + 714E^2P_2^2R^2$$
$$+70E^5P_1P_0R^2 + 182E^4P_2P_0R^2 + 1050E^2P_1P_2R + 420E^3P_0P_2R$$
$$+1575EP_2^2R + 210E^4P_0P_1R + 210E^3P_1^2R + 70E^5P_0^2R + 420E^2P_0P_2$$
$$+1050P_1P_2E + 210P_0P_1E^3 + 70P_0^2E^4 + 210E^2P_1^2 + 1575P_2^2\ )$$

Figure 4. Explicit two-center integrals for a radial-function prototype, using the radial functions in eq. 22 (which for simplicity have not been normalized).

for Si_3 that is a few percent too large, and a bond length for Si_4 that is a few percent too small, several *other* local minima will give very similar results. This is a remarkable observation when one considers that we use 200 or more such bond lengths, binding energies, etc. and that we observe very strong correlation of different local minima across all 200 properties. Also, this behavior seems to be widespread, at least in our own experience, as we have observed this for just about every non-trivial parameter fitting that we have done.

This strong correlation of different local minima suggests that, loosely speaking, these "different" local minima are really not different at all, but rather in some way they represent *the same* local minimum. This would mean that the local minima are connected by hidden constraints, and that the problem would be less poorly behaved if one accounted for these constraints. This can be better understood if one considers the situation in reverse: start with a well-behaved function with a small number of inherently different local minima. Then maliciously introduce some spurious and highly non-linear degrees of freedom into the function. What would happen? One might expect this to wreak havoc on the function, causing just the type of behavior that we observe, as well-defined local minima split or bifurcate with the introduction of spurious dimensions. As we have not significantly tested our radial function prototype, we can not claim that it is the integrability constraints that are responsible for the poor behavior of the optimization function. However, if they are responsible, then treating $S_{ss\sigma}(R)$ etc. as independent functions would have a devastating effect on the ability to find the global minimum. This would mean that a parameter fitting calculation would waste large amounts of time exploring the spurious dimensions introduced by removing the constraints.

E Preliminary results

Although still a prototype, we have performed an initial test that supports a radial function model. We have applied our environment dependent model, which consists of a two-center model consisting of the function $S_{ss\sigma}(R)$, $H_{ss\sigma}(R)$, etc., and also consists of environment dependent modifications discussed elsewhere in this report, to the single-element systems of C, Si, and Ge. For logistical reasons C and Si were more heavily optimized, to the point that we now have stable sets of parameters for these two elements; Ge also has been very successfully optimized. The work on C and Si consisted not only of parameter optimization in the direct sense, but also of subsequent testing of several "candidate" sets of parameters that were eventually discarded in favor of one "final" set.

For this initial test, we used our results for the overlap functions $S_{ss\sigma}(R)$, $S_{sp\sigma}(R)$, $S_{pp\sigma}(R)$, $S_{pp\pi}(R)$ to study our deconvolution argument, i.e. that the existence of overlap functions implies the existence of radial functions. In the work leading to these sets of parameters for $S_{ss\sigma}(R)$, etc., we occasionally used parameter constraints on the overlap functions. However, for the most part, these four function we parameterized *independently of each other.* The results of this test suggest that even though the *four* overlap functions were parameterized independently, they can be approximately obtained from only *two* radial functions.

For this test, we performed a least squares curve fitting of our four optimized overlap functions to the same four functions obtained from analytical integration of the following radial functions:

$$\begin{aligned} \phi_s(r) &= (S_0 + S_1 \cdot r + S_2 \cdot r^2) \cdot exp(-E \cdot r) \\ \phi_p(r) &= (P_0 + P_1 \cdot r + P_2 \cdot r^2) \cdot exp(-E \cdot r) \end{aligned} \tag{23}$$

This relatively simple form was chosen for simplicity; in general one could use higher-order polynomial coefficients and, in particular, could use different values for the exponents. The curve fitting was performed by first extracting $\phi_s(r)$ from

$S_{ss\sigma}(R)$, and extracting $\phi_p(r)$ from $S_{pp\sigma}(R)$, then by constructing $S_{sp\sigma}(R)$ and $S_{pp\pi}(R)$ from $\phi_s(r)$ and $\phi_p(r)$. The results are shown in Figure 5.

Even with this overly-simplified form for $\phi_s(r)$ and $\phi_p(r)$, these results suggest that the large-scale empirical fitting process is driving the system toward results that are consistent with the existence of not four but only two independent functions, i.e. $\phi_s(r)$ and $\phi_p(r)$. This is an important result in light of the fact that our final sets of parameters for C and Si involved a large amount of computational resources. During the course of the optimization, we encountered the problematic nature of the optimization problem discussed earlier. This includes: (1) the optimization consisted of large number poorly-distributed local minima, resulting in the need for a greatly increased time to find the global minimum, (2) the overlap matrix was frequently not positive definite, resulting in sets of parameters for which systems of atoms did not have a calculatable energy (and also interrupting the optimization algorithm), (3) several different local minima gave very similar results to each other, suggesting the existence of hidden constraints, (4) parameters with reasonable calculated values but unreasonable parameter values, such as long range $S_{ss\sigma}(R)$, also suggesting the existence of hidden constraints, and (5) the failure of candidate sets of parameters in subsequent testing, also suggesting the existence of hidden constraints.

In conclusion, this initial test suggests that the treatment of $S_{ss\sigma}(R)$, $S_{sp\sigma}(R)$, $S_{pp\sigma}(R)$, $S_{pp\pi}(R)$ as independent functions might have a devastating effect on the global optimization problem, i.e. by requiring an excessive amount of computational resources to find the global minimum. In the language of global optimization, we have found approximate constraints that greatly reduce the parameter space that one must search for the global minimum in. Finally, it perhaps is remarkable that the individual deconvolutions exist, i.e. that starting with $S_{ss\sigma}(R)$, one can obtain $\phi_s(r)$ that reproduces $S_{ss\sigma}(R)$. This is remarkable because the existence of a solution to a deconvolution problem can not in general be

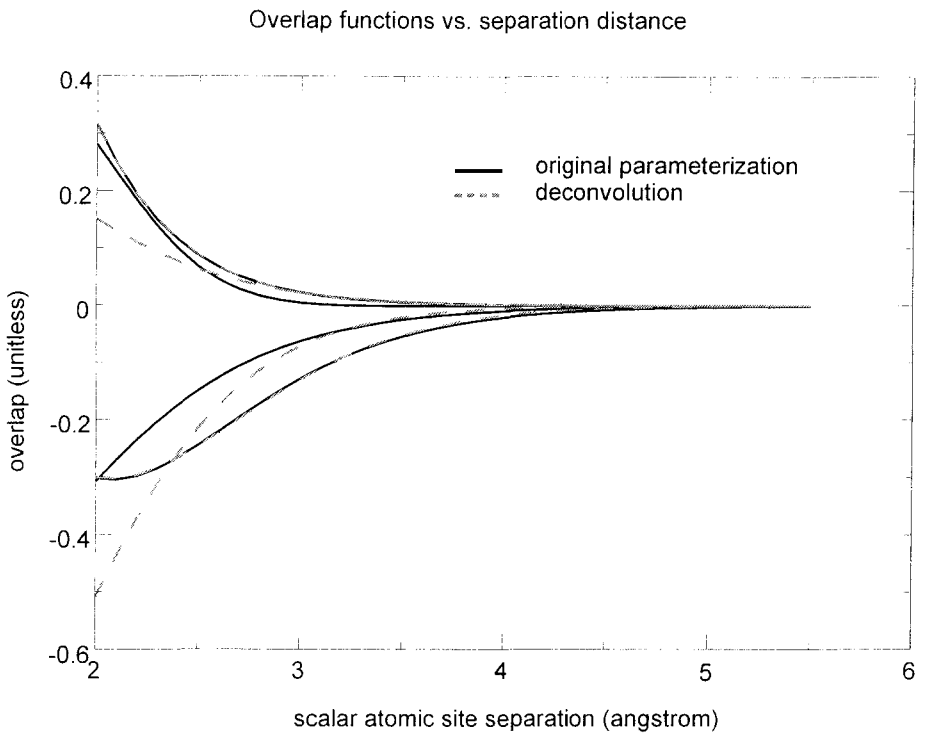

Figure 5. Preliminary test of our radial function prototype, using the radial functions in eq. 23. The overlap functions used in our current model are shown (solid), together with the same functions obtained from a deconvolution procedure (dashed).

guaranteed. Our initial test suggests that this deconvolution is well-defined as one considers the overlap integrals individually.

CHAPTER IV

RADIAL FUNCTION PROTOTYPE

A Nitrogen and Oxygen

Of the many classifications of materials science models that can be made, one is to distinguish models that have their origins in the semiconductor community from those that have their origins in the organic materials community. For models in the first group, including our own, the original idea of treating multi-element systems was of course to treat systems of interest to the semiconductor community. However, with the increasing demand for organic and biological applications it is perhaps these systems that are now more interesting. Although the technical differences between the two categories are subtle, models in the first category are rarely used to study organic systems other than simple hydrocarbons. This makes the native treatment of nitrogen and oxygen by our radial function prototype particularly interesting. While we are on the subject of nitrogen and oxygen, we should not overlook the important industrial and military applications for these elements. Transition metal oxides are a representative example of important applications that are outside both the semiconductor and the organic communities.

We have already argued in Section A that models based on the based on the 9 functions in eq. 1 do not treat multi-element averaging natively. In this section we are going to argue that, even if for the sake of argument they did, that they still do not treat nitrogen and oxygen natively. The problem with these elements is that, apart from any problems caused by multi-element averaging, tight-binding models *can not treat either nitrogen or oxygen separately*, i.e. as single-element-only

systems. This is caused by the fact that, as individual elements, nitrogen and oxygen do not form enough structures from which one can obtain a list of reference properties needed for empirical fitting. As single-element systems, these two elements do little more than form the dimers N_2 and O_2. Nitrogen and oxygen do have crystalline structures, but these consist only of isolated dimers weakly bonded to each other by van der Waals forces.

We should point out that in hindsight it is perhaps something of a coincidence that elements such as C and Si can be treated. The difference is actually quite subtle, as C and Si as single elements do not readily form a large number of structures suitable for fitting either. The band structure of diamond Si, along with a few other experimentally observed structures such as Si_2, is not enough for a large-scale fitting. The subtle difference is that although they do not *readily* form, there are still a large number of structures that can be studied using first-principles calculations. Beginning in the 1980s, most notably with the work of Cohen for crystalline Si [19] and the work of Raghavachari for C and Si clusters [20], there arose a very loose standard of computationally well-defined crystalline phases and small clusters, making such large-scale fitting possible. This includes at least six different crystalline phases and at least 20 different small clusters for Si. Unfortunately, nitrogen and oxygen do not form enough structures suitable for fitting that can be studied either experimentally or with first-principles calculations.

Let us then return to the tight-binding averaging scheme for multi-element systems in eq. 14:

$$S_{ss\sigma}^{\mathrm{C-N}}(R) = \tfrac{1}{2} \cdot \left(S_{ss\sigma}^{\mathrm{C-C}}(R) + S_{ss\sigma}^{\mathrm{N-N}}(R) \right)$$

We can now see that, even if for the sake of argument we assume that this type of averaging works, it still can not treat nitrogen and oxygen. The averaging is based on the existence of a parameter fitting for the single-element systems, but for

nitrogen and oxygen we do not *have* such a parameter fitting:

$$S_{ss\sigma}^{\mathrm{C-N}}(R) = \tfrac{1}{2} \cdot \left(S_{ss\sigma}^{\mathrm{C-C}}(R) + (???) \right)$$

Our radial function prototype avoids these problems by constructing the two-center integrals directly, without any reference to N-N or O-O:

$$S_{ss\sigma}^{\mathrm{C-N}}(R) = \int_{\mathbf{r}} \phi_s^{\mathrm{C}} \left(\mathbf{r} - \mathbf{R}_i \right) \cdot \phi_s^{\mathrm{N}} \left(\mathbf{r} - \mathbf{R}_j \right) \cdot d\mathbf{r}$$

One can then completely avoid any need to treat nitrogen and oxygen separately, and can proceed to treat multi-element systems directly or natively.

To be fair, we should mention that there are some creative ways of working around this problem within tight-binding. One option is to maintain the parameterization of the N-N or O-O functions, but to fit only to multi-element systems, i.e. by using some averaging scheme to construct the multi-element interactions. This is a questionable technique that does not appear to be widely used. The next and most widely used option is to avoid any averaging scheme, and to parameterize functions such as $S_{ss\sigma}^{\mathrm{C-N}}(R)$ individually. If we overlook the issue of positive definite overlap, this is actually a reasonable technique for problems consisting of only two different elements. However, for more complex systems one again encounters the explosion of parameters discussed in Section A. Here also, we have the paradox that the earlier or simpler models are capable of treating nitrogen and oxygen natively, while it is with the more complicated or more general models that things start to break down.

Finally, we should point out that a radial function prototype raises the possibility of fitting "across the boards", i.e. where the radial functions of several elements are fit at the same time. For organic and biological systems, one could start with $\{\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}\}$ and then fit to a large list of small organic molecules. This is of course not a new idea, as it is exactly the type of fitting preferred by the models used in the organic community. However, such a fitting has never been attempted by any of the models that we are in competition with. A particularly appealing

feature of fitting across the boards is that one can fit to experimentally known values, as it is perhaps a legitimate criticism of existing models that they rely too heavily on first-principles calculations, particularly density functional theory. In hindsight this criticism might also be applied to Si itself and to other semiconductor elements, i.e. apart from any of the complications caused by nitrogen and oxygen.

B Not-just-energy properties

The empirical orbital models that are the subject of this report, including our radial function prototype, and including models with environment-dependent modifications, all use a matrix form of the Schrödinger equation:

$$H_{i\alpha,j\beta} \cdot C_{j\beta,\lambda} = S_{i\alpha,j\beta} \cdot C_{j\beta,\lambda} \cdot E_{\lambda}$$

where H and S are input, and C and E are output. Although not explicitly part of the matrix eigenvalue equation, the one-electron wave functions Ψ_{λ} are also implied as part of the output:

$$\Psi_{\lambda}(\mathbf{r}) = \sum_{i\alpha} C_{i\alpha,\lambda} \cdot \Phi_{i\alpha}(\mathbf{r})$$

The wave functions Ψ_{λ} "inherit" their dependence on the position $\mathbf{r}$ from the atomic orbitals Φ, as indicated in this expression. It is with this position dependence that existing models are problematic. By directly parameterizing the functions $S_{ss\sigma}(R)$ etc., such models never explicitly specify the atomic orbitals, leaving the position dependence of the wave functions undefined:

$$\Psi_{\lambda}(\mathbf{r}) = \sum_{i\alpha} C_{i\alpha,\lambda} \cdot (???)$$

With a radial function prototype, the atomic orbitals are parameterized directly, and the position dependence of Ψ_{λ} is restored.

To be fair, we should acknowledge that existing models do treat the calculation of *some* properties other than the energy natively. In fact, it is the existence of electronic structure information that distinguishes first-principles and

empirical orbital models from molecular mechanics and finite element models. The limitation is that existing models can only obtain not-just-energy properties from objects that are present in the eigenvalue equation. Properties that can not be expressed in terms of these objects can not be evaluated. The charge density $\rho(\mathbf{r})$ is probably the most important example:

$$\rho(\mathbf{r}) = \sum_{\lambda} N_{\lambda} \cdot \Psi_{\lambda}(\mathbf{r}) \cdot \Psi_{\lambda}(\mathbf{r})$$

Contour plots of $\rho(\mathbf{r})$ are the most widely used tool to visualize electronic structure information. In practice one can obtain such plots for tight-binding models by introducing some "characteristic set" of atomic orbitals. In light of our discussion, such characteristic orbitals look very much like attempts (i.e. poor attempts) to obtain radial functions as deconvolutions of the two-center integrals.

The charge density $\rho(\mathbf{r})$ and other functions of $\mathbf{r}$ are involved in a very wide variety of electronic structure applications. However, the problematic nature of not-just-energy properties is perhaps better understood by considering properties that are *not* functions of $\mathbf{r}$. To understand this in more detail, we will consider a representative example, that of *atomic polar tensor* charges. The starting point of this discussion is the need to calculate the charge "associated with" individual atoms, which is closely related to the more qualitative concepts of ionic and covalent bonding. When position information is not available, one attempts to construct an expression for the charge using objects that are present in the eigenvalue equation. The most straightforward approach leads to an expression for the total number of electrons N_{total}:

$$\sum_{\lambda, i\alpha, j\beta} N_{\lambda} \cdot C^{*}_{i\alpha,\lambda} \cdot S_{i\alpha,j\beta} \cdot C_{j\beta,\lambda} = \sum_{\lambda} N_{\lambda} = N_{total}$$

By removing some of the summations over $i\alpha$ and $j\beta$, one can identify charges associated with various combinations of sites and orbitals. Charges obtained from this expression are usually called *Mulliken charges*.

The Mulliken analysis is not a particularly bad way of calculating atomic charges. However, it is not a particularly good way either. Amid concerns over their accuracy, several first-principles calculations have replaced these types of charge analyses with more elaborate ones. In the atomic polar tensor analysis, which is attributed to Cioslowski [21], charges are obtained from derivatives of the dipole moment:

$$Q_i = \frac{1}{3} \cdot \left(\frac{\partial}{\partial X_i} \mu_x + \frac{\partial}{\partial Y_i} \mu_y + \frac{\partial}{\partial Z_i} \mu_z \right)$$

with:

$$\mu_x = \sum_{\lambda, i\alpha, j\beta} N_\lambda \cdot C^*_{i\alpha,\lambda} \cdot C_{j\beta,\lambda} \cdot \int_{\mathbf{r}} \Phi_{i\alpha} \left(\mathbf{r} - \mathbf{R}_i \right) \cdot x \cdot \Phi_{j\beta} \left(\mathbf{r} - \mathbf{R}_j \right) \cdot d\mathbf{r} \qquad (24)$$

For the purposes of our discussion in this section, we are not interested in debating the accuracy of the Mulliken analysis. It is as a representative example that the Mulliken vs. APT debate reveals the limitations of how existing models treat not-just-energy properties. Here we have two different models or "analyses" for the concept of atomic charges. The Mulliken analysis, in hindsight perhaps by coincidence, can be expressed entirely in terms of objects that are present in the eigenvalue equation. The APT analysis can not, even though it represents the same concept of atomic charge.

If we look more closely at eq. 24, we can see where things start to go wrong. Existing models are only aware of position $\mathbf{r}$ integrals if they happen to be the overlap $\int \Phi_i \cdot \Phi_j$ or the Hamiltonian.$\int \Phi_i \cdot \hat{\mathrm{H}} \cdot \Phi_j$. If the integral is changed slightly, in this case to $\int \Phi_i \cdot x \cdot \Phi_j$, existing models break down. It is this special ability to "look inside" the integrals that gives a radial function prototype the advantage. The only evident way to treat such properties within the existing framework would be to introduce more parameterized functions. This brings us back to our deconvolution argument, as independently parameterized functions would imply multi-valued radial functions. Again we have the interesting conclusion that our prototype is in some ways a very complicated set of parameter constraints. That is, one can

perform the integrations "on paper", and them use them in calculations, as if the atomic orbitals did not exist. The APT example shows that this concept of constraints applies not only to the overlap and Hamiltonian functions, but also to integrals such as $\int \Phi_i \cdot x \cdot \Phi_j$.

C Hamiltonian orbitals

Our radial function prototype treats multi-element systems natively, at least as far as the overlap matrix is concerned. However, to completely satisfy the multi-element requirement, the Hamiltonian must also treat multi-element systems natively. To develop a prototype for the Hamiltonian, we return to our discussion in Section D, where the matrix elements of the Hamiltonian consist of the terms $H_{ij,\nabla}$ and $H_{ij,k}$:

$$H_{ij,\nabla} = \int_{\mathbf{r}} \Phi_i (\mathbf{r} - \mathbf{R}_i) \cdot \nabla_{\mathbf{r}}^2 \cdot \Phi_j (\mathbf{r} - \mathbf{R}_j) \cdot d\mathbf{r}$$

$$H_{ij,k} = \int_{\mathbf{r}} \Phi_i (\mathbf{r} - \mathbf{R}_i) \cdot V_k (\|\mathbf{r} - \mathbf{R}_k\|) \cdot \Phi_j (\mathbf{r} - \mathbf{R}_j) \cdot d\mathbf{r}$$

It is again important to point out that although several layers of approximation are needed to obtain this form for H, these layers of approximation are already required in order to construct the two-center integrals $H_{ss\sigma}(R)$, $H_{sp\sigma}(R)$, $H_{pp\sigma}(R)$, $H_{pp\pi}(R)$. Within the context of this report, it is the existence of $H_{ss\sigma}(R)$ etc. that defines a "two-center model", and we will take this as a starting point for our discussion. This is not an arbitrary or semantic definition, as there are about six research groups that we are in competition with, and each of them use a $H_{ss\sigma}(R)$ etc. framework for their environment-dependent model.

As a first pass at a prototype for the Hamiltonian, it is evident that if one already has explicit functional forms for the radial functions $\phi_s(r)$ and $\phi_p(r)$, then one can construct the terms $H_{ij,\nabla}$ without any additional empirical parameters, i.e. by operating on the orbitals $\Phi(\mathbf{r})$ explicitly with $\nabla_{\mathbf{r}}^2$. It is also evident that one can construct the terms $H_{ij,k}$ if one introduces a parameterized function $V(r)$ for the

potential. Note that within the central field approximation, the three-dimensional potential $V(\mathbf{r})$ of an arbitrary configuration of atoms is completely specified by a one-dimensional scalar function $V(r)$ at each atomic nucleus, which is presumably the same for each type of element. This leads naturally to a treatment of multi-element systems, as the operator $\nabla^2_{\mathbf{r}}$ is the same for each element, and as each potential term V_k is associated with a specific element, i.e. the element of the nucleus located at $\mathbf{R}_k$:

$$H_{ij,\nabla} \overset{\text{first pass}}{=} \int_{\mathbf{r}} \Phi_i^{elem(i)}(\mathbf{r}-\mathbf{R}_i) \cdot \nabla^2_{\mathbf{r}} \cdot \Phi_j^{elem(j)}(\mathbf{r}-\mathbf{R}_j) \cdot d\mathbf{r}$$

$$H_{ij,k} \overset{\text{first pass}}{=} \int_{\mathbf{r}} \Phi_i^{elem(i)}(\mathbf{r}-\mathbf{R}_i) \cdot V_k^{elem(k)}(\|\mathbf{r}-\mathbf{R}_k\|) \cdot \Phi_j^{elem(j)}(\mathbf{r}-\mathbf{R}_j) \cdot d\mathbf{r} \qquad (25)$$

where $elem(i)$ refers to the identity of the element indexed by i. It is important to point out that the prefactor $\frac{-\hbar^2}{2m}$ which has been incorporated into the operator $\nabla^2_{\mathbf{r}}$ involves the mass of the electron, which is the same for each element; the prefactor $\frac{-\hbar^2}{2M}$ involving the mass of the nucleus in involved in the subsequent motion of the atoms, but is not involved in the eigenvalue equation.

This first pass at a prototype for the Hamiltonian is problematic for two reasons. First, it is widely agreed that even highly accurate first-principles calculations usually do not give accurate values for the atomic orbitals $\Phi(\mathbf{r})$ themselves, i.e. as values of the probability density of an individual electron. Operating on $\Phi(\mathbf{r})$ with $\nabla^2_{\mathbf{r}}$ means that the Hamiltonian matrix elements will consist of *second-order differences* of the values of $\Phi(\mathbf{r})$ at adjacent values of $\mathbf{r}$. This suggests that it is not appropriate to operate on parameterized radial functions with $\nabla^2_{\mathbf{r}}$. The second problem is that the resulting model places too much emphasis on the overlap and not enough on the Hamiltonian. The real Hamiltonian operator is very complicated, and there is a general feeling that one should have at least as many parameters for the Hamiltonian as for the overlap. This is the case with the 9 functions in eq. 1, where there are 4 degrees of freedom each for the overlap and Hamiltonian. However, our first pass at a prototype results in two or three degrees

of freedom $\phi_s(r)$, $\phi_p(r)$, $\phi_d(r)$ for the overlap but only one degree of freedom $V(r)$ for the Hamiltonian.

The crux of the matter is that is never necessary to specify an explicit form for the operator $\nabla^2_{\mathbf{r}}$, i.e. one can replace $\nabla^2_{\mathbf{r}}$ with a more general unknown or unspecified operator, as long as certain conditions are satisfied. The most important of these conditions is that the operator modifies only the radial part of an atomic orbital and not the angular part:

$$\nabla^2_{\mathbf{r}} \cdot \phi(\theta,\varphi) \cdot \phi(r) = \phi(\theta,\varphi) \cdot \nabla^2_r \cdot \phi(r) = \phi(\theta,\varphi) \cdot \phi^{\nabla}(r)$$

with:

$$\phi^{\nabla}(r) = \nabla^2_r \cdot \phi(r)$$

Because of this condition, the Hamiltonian integrals have the same symmetry properties as the overlap integrals; it is this condition that allows one to construct $H_{ss\sigma}(R)$ etc.. Our argument is that we can make a second pass at a prototype for the Hamiltonian by treating the functions $\phi^{\nabla}(r)$ as degrees of freedom:

$$\phi^{\nabla}_s(r) = \left\{\nabla^2_r\right\} \cdot \phi_s(r)$$
$$\phi^{\nabla}_p(r) = \left\{\nabla^2_r\right\} \cdot \phi_p(r)$$

where $\left\{\nabla^2_r\right\}$ is some generalization of the ∇^2_r operator. This second pass results in a model with two or three degrees of freedom $\phi_s(r)$, $\phi_p(r)$, $\phi_d(r)$ for the overlap, and three or four degrees of freedom $\phi^{\nabla}_s(r)$, $\phi^{\nabla}_p(r)$, $\phi^{\nabla}_d(r)$, $V(r)$ for the Hamiltonian.

This second pass still satisfies our requirement for multi-element systems. The "∇^2 orbitals" $\phi^{\nabla}(r)$ have a well-defined association with a specific element, because $\phi(r)$ is associated with a specific element, and because $\left\{\nabla^2_r\right\}$ is not associated with any element. Two additional comments are also in order. First, with explicit functional forms for $\phi(r)$ and $V(r)$, it is possible to explicitly evaluate the *three-center integrals*. Now, we are not suggesting that such integrals be included in an empirical orbital model. From a performance standpoint, three-center integrals are very costly and would slow down a calculation to an unacceptable level. There

also appears to be a growing consensus that the most important extension to a two-center framework involves the electron density $\rho(\mathbf{r})$ rather than other extensions such as three-center integrals. It is evident however, that the ability to calculate such integrals could be quite useful for testing and reference purposes.

The second comment is that the treatment of $\left\{\nabla_r^2\right\}$ as a generalization of the ∇_r^2 operator is quite consistent with existing overlap-parameterized models. This can be illustrated by returning to our deconvolution argument. Here, we will dispense with any complications due to multi-valued radial functions and consider only the $ss\sigma$ and $pp\sigma$ interactions. For the overlap, we have two degrees of freedom $S_{ss\sigma}(R)$ and $S_{pp\sigma}(R)$ before the deconvolution, and two degrees of freedom $\phi_s(r)$ and $\phi_p(r)$ after. However, if $\phi_s^\nabla(r)$ and $\phi_p^\nabla(r)$ are not treated as a degrees of freedom, then for the Hamiltonian we have two degrees of freedom $H_{ss\sigma}(R)$ and $H_{pp\sigma}(R)$ before the deconvolution, but only one degree of freedom $V(r)$ after. If we temporarily dispense with the complications caused by $V(r)$, then this is consistent with a treatment of $\phi_s^\nabla(r)$ and $\phi_p^\nabla(r)$ as degrees of freedom, rather than as resulting from some specific operator ∇_r^2. One can go so far as to obtain expressions for $\phi_s^\nabla(r)$ and $\phi_p^\nabla(r)$ as functionals:

$$\phi_s^\nabla(r) \overset{\text{deconv.}}{=} \phi_s^\nabla(r)\left[S_{ss\sigma}(R), H_{ss\sigma}(R)\right]$$
$$\phi_p^\nabla(r) \overset{\text{deconv.}}{=} \phi_p^\nabla(r)\left[S_{pp\sigma}(R), H_{pp\sigma}(R)\right]$$

Of course, with the potential $V(r)$ this strict one-to-one correspondence of degrees of freedom breaks down. Still, one can make a strong argument that tight-binding models, i.e. models that treat $S_{ss\sigma}(R)$, $H_{ss\sigma}(R)$, etc. as degrees of freedom, imply the existence of one or more degrees of freedom associated with some generalization of the ∇_r^2 operator.

Finally, one can make a third pass at the Hamiltonian by treating the potential $V(r)$ not as a function but as an operator. This leads to "potential orbitals" $\phi_s^V(r)$ and $\phi_p^V(r)$:

$$\phi^V(r) = \hat{\mathrm{V}}_r \cdot \phi(r)$$

The potentials $V(r)$ and the radial functions $\phi(r)$ are both associated with a specific element, and so the resulting model will not treat multi-element systems natively unless $V(r)$ and $\phi(r)$ always index the same element. Now, one can argue that within the two-center approximation, the integrals $\int \Phi_i \cdot V_k \cdot \Phi_j$ are taken to be zero unless $k = i$ or $k = j$. In this case one will always have either $\int (\Phi_i \cdot V_i) \cdot \Phi_j$ or $\int \Phi_i \cdot (V_j \cdot \Phi_j)$, and then $\phi^V(r)$ can be associated with a specific element. This argument, however, does not account for the "neglected two-center integrals", which will be discussed later in this chapter. For now we can point out that this third pass will not lead to two degrees of freedom $\phi^\nabla(r)$ and $\phi^V(r)$ for each type of orbital, because the off-site matrix elements of the Hamiltonian will always factor as:

$$H_{ij}|_{i \neq j} = \int \Phi_i \cdot \left(\tfrac{1}{2}\Phi_j^\nabla + \Phi_j^V\right) + \int \left(\tfrac{1}{2}\Phi_i^\nabla + \Phi_i^V\right) \cdot \Phi_j \tag{26}$$

As long as the neglected two-center integrals are treated separately, this leads to a reasonable model with two or three degrees of freedom $\phi_s(r)$, $\phi_p(r)$, $\phi_d(r)$ for the overlap, and two or three degrees of freedom $\phi_s^H(r)$, $\phi_p^H(r)$, $\phi_d^H(r)$ for the Hamiltonian, with $\phi^H(r) = \frac{1}{2}\phi^\nabla(r) + \phi^V(r)$, although it will not be possible to identify $\phi^\nabla(r)$ and $\phi^V(r)$ individually.

D Neglected two-center integrals

We have taken a policy-driven approach to our discussion of the next generation of two-center models, i.e. an approach that starts with a list of requirements and then seeks models that satisfy those requirements. Now, the concept of a *two-center model* is that matrix elements consist of integrals involving one, two, three, or (in some cases) four centers or atomic nuclei, and that a reasonable model can be obtained by treating only the one and two center integrals. Within the central field approximation, there is a very well-defined accounting of the total number of integrals: N^2 for the overlap, N^2 for the ∇^2 terms of the Hamiltonian, and N^3 for the V terms of the Hamiltonian. Of these N^3 terms, it is

an elementary matter of combinatorics that there are $N \cdot (N-1) \cdot (N-2)$ terms with three distinct centers, leaving $3N^2 - 2N$ terms that contain either one or two distinct centers. From a policy-driven approach it then seems obvious that a two-center model should be required to treat not just *some* of the two-center integrals, but *all* of the two-center integrals.

Existing tight-binding models do *not* satisfy this requirement. For the V terms there are three indexes i, k, j. Two distinct centers are obtained for $i = k \neq j$ and for $i \neq k = j$. However, two distinct centers are *also* obtained for $k \neq i = j$. These "neglected" two-center integrals consist of orbitals i and j which are located at the same nucleus and a potential term k located at a different nucleus. Following the notation of Section D and E, if these integrals are accounted for then the on-site matrix elements of the Hamiltonian have the form:

$$H_{ii} = \varepsilon_i + \textstyle\sum_{k \neq i} H_{ii,k} \tag{27}$$

Although it would be quite possible to include these terms in a tight-binding model, almost all existing models treat the on-site elements as fixed constants, which corresponds to taking $H_{ii,k} = 0$. However, unlike three-center integrals, which one can argue are relatively small, there is no reason to believe that these integrals are any smaller than other two-center integrals that are not neglected. We will then add a final requirement to our list: the model *must treat the neglected two-center integrals.*

As with other topics in this chapter, this problem can be better understood by considering the historical development of tight-binding models. Early models started with periodic crystalline structures as the fundamental type of material, the most important of which by far was diamond Si. The high symmetry of such simple periodic structures often results in a cancellation of quantities that do not otherwise cancel. Similarly, the existence of a well-defined coordination number often results in "effective constants" for interactions that are otherwise more complicated. As discussed in an article by Mercer and Chou [22], this is indeed the case with the

neglected two-center integrals. For example, the summation in eq. 27 is zero for the s, p_x and p_x, p_y interactions for several types of crystal symmetries, including the cubic symmetry of diamond Si. For simple crystal structures, the remaining interactions result in effective constants which have only a simple dependence on coordination number and coordination distance.

From a historical perspective then, the use of tight-binding models for complicated non-periodic systems is a fairly recent development. When simple periodic structures were the fundamental type of material, it was probably appropriate to incorporate the neglected two-center integrals into the on-site energies ε_s and ε_p. However, the resulting "standard model" of $S_{ss\sigma}(R)$, $H_{ss\sigma}(R)$, etc. was then carried over to the modern arena, where complicated non-periodic structures are now the fundamental type of material. It is also quite likely that the lack of treating these integrals is in some cases a simple mistake or oversight. It is very easy to take the combination $\Phi_i \cdot V_k \cdot \Phi_j$ and then make a "two-center approximation" that retains only those combinations with $k = i$ or $k = j$. This oversight is suggested by the fact that journal articles on tight-binding models rarely mention these integrals or offer any explanation of why they are excluded from a model. In fact, it is not clear to what extent the on-site energies obtained by existing models contain the "coordination constant" effects of the neglected integrals. This could lead to undesirable behavior, as the energies ε are the energies of the isolated atom, which of course should not have any coordination energy.

Both our first and second passes at the Hamiltonian in Section C are already capable of meeting our new requirement. This simply involves explicitly evaluating the integrals $\int \Phi_i \cdot V_k \cdot \Phi_i$ using the overlap functions $\phi(r)$ and the potential function $V(r)$. Multi-element systems are treated natively; a detailed expression for the integration with element identities has already been given in eq. 25. Using these integrals in large-scale calculations does not appear to be problematic. Following the notation of Mercer and Chou [22], one sets up the scalar functions $I_{ss\sigma}(R)$,

$I_{sp\sigma}(R)$, $I_{pp\sigma}(R)$, $I_{pp\pi}(R)$. In fact, the neglected two-center integrals apparently satisfy exactly the same transformation relations as the overlap and Hamiltonian integrals, for example (see eq. 21):

$$I_{x,x} = l^2 \cdot I_{pp\sigma} - (1 - l^2) \cdot I_{pp\pi}$$

The performance cost is also reasonable. Of the $3N^2 - 2N$ terms for the potential, there are only $N \cdot (N - 1)$ neglected terms. Other previous items of discussion, such as the possibility of obtaining analytical forms for $I_{ss\sigma}(R)$ etc., also apply here.

It is important to point out that, following our second pass in Section C, the parameterization of $V(r)$ is meaningful only if it is used in these neglected integrals. Otherwise, the factorization of the off-site elements in eq. 26 will result in the collapse of $V(r)$ as a degree of freedom. That is, without the neglected integrals, the Hamiltonian can be expressed entirely in terms of the Hamiltonian orbitals $\phi^H(r)$, without any reference to $V(r)$. It is only by using $V(r)$ in both the on-site and off-site elements that complete factorization does not occur, making it possible to treat $V(r)$ as a parameterized function. One might want to make a third pass at the Hamiltonian, as in Section C. In this case however, it does not appear to be meaningful to introduce "potential orbitals", i.e. to replace the potential function with an operator. The multiplication $V_k \cdot \Phi_i$ does not result in an orbital centered at $\mathbf{R}_i$, and even if it did the resulting radial function $\phi^V(r)$ would be associated with two different elements.

There still does appear to be at least one meaningful generalization, and that is to use a different potential function $V^I(r)$ for the neglected integrals than for the other Hamiltonian integrals. Due to the factorization in eq. 26, this does not result in a net increase in the number of parameters. However, the resulting Hamiltonian, with degrees of freedom $\phi_s^H(r)$, $\phi_p^H(r)$, $\phi_d^H(r)$, and $V^I(r)$, is perhaps the most appealing of all choices. First, it achieves a better balance between the ∇^2 and V terms. Just as one expects the Hamiltonian to be more complicated than the overlap, one expects the potential to be more complicated than the kinetic energy.

Our second pass in Section C can be criticized for placing too much emphasis on the kinetic energy. Next, this final pass is consistent with the possibility that $V(r)$ and $V^I(r)$ are associated with fundamentally different types of chemistry, a possibility suggested by the success of models that do not treat the neglected integrals. Finally, the decoupling of $V^I(r)$ from the off-site terms is expected to assist the optimization algorithm, in that a change in the parameters of $V^I(r)$ will not change the Hamiltonian orbitals $\phi^H(r) = \frac{1}{2}\phi^\nabla(r) + \phi^V(r)$.

A final note that one should use caution when implementing these neglected integrals, as the variety of options that we have enumerated can lead to some confusion. Most importantly, we should clarify the various options for treating the one-center integrals $\int \Phi_i \cdot V_i \cdot \Phi_i$. If one treats each Hamiltonian integral individually, the one-center integrals $H_{ii,i}$ appear explicitly in the on-site elements:

$$H_{ii} = H_{ii,\nabla} + H_{ii,i} + \sum_{k \neq i} H_{ii,k}$$

However, if one starts with the on-site energies ε_i, the one-center integrals do not appear explicitly, as they are already incorporated in ε_i:

$$H_{ii} = \varepsilon_i + \sum_{k \neq i} H_{ii,k}$$

Finally, if one avoids the on-site energies ε_i and instead starts with the off-site elements and their limiting values ε_i', the one-center integrals appear explicitly, but with a minus sign:

$$H_{ii} = \varepsilon_i' - H_{ii,i} + \sum_{k \neq i} H_{ii,k}$$

The minus sign results from the fact that each off-site value ε_i' implicitly contains two occurrences of $H_{ii,i}$. Note also that the factorization in eq. 26 makes it easy to make a "Murphy's law" mistake of being off by a factor of two. The traditional tight-binding functions $H_{ss\sigma}$ etc. each account for one kinetic and two potential terms, while our Hamiltonian orbitals ϕ^H each account for one-half kinetic and one potential term.

CHAPTER V

OPTIMIZATION TECHNIQUES

A "Given a function"

"Given a function $F(X)$, find the value $X_{\min}$ such that $F(X_{\min}) < F(X)$ for all other values of X." This is the widely-encountered *minimization* problem. If we had an analytical form for $F(X)$, it might be very easy to find $X_{\min}$. However, for the numerical problem the only information that we can ever know about $F(X)$ is the specific numerical value of F, and in some cases the numerical values of the derivatives of F, for a specific numerical value of X. This means that even the most sophisticated minimization algorithm will need to send one value of X to the function F, then another value of X, then another, until the algorithm is reasonably certain that it has found the minimum value. The difficulty with minimization problems is that the time required to find the minimum value can vary over many orders of magnitude depending on the problem. In some cases the problem might be unsolvable, even on the fastest computer.

A closely related problem is the *root-finding* problem: "Given a function $F(X)$, find the value X_{root} such that $F(X_{root}) = 0$." Of course, if something other than zero is on the right-hand side of this equation, the equation can always be expressed as $F(X_{root}) - G(X_{root}) = 0$, which is still a root-finding problem. In many applied problems the function F depends on several values $\{X_1, X_2, X_3...\}$, and one has a *multi-dimensional* problem. For the multi-dimensional minimization problem one still has a scalar value for F, but for the multi-dimensional root-finding problem the existence of a well-defined solution requires that F and X have the same

dimension. Finally, for the multi-dimensional minimization problem, if the function F can be expressed as a sum of squared terms, one has a *least-squares* problem. Although it might seem that this should be treated as any other minimization problem, the structure of a least-squares function can be exploited to find the minimum in much less time than for a general function. Of course, there must be a relatively large number of terms in the least-squares sum: in order to exploit least-squares algorithms there must be at least as many terms in the sum as there are dimensions in X.

The minimization problem is to find the *global* minimum, which is the set $\{X_i\}$ with the absolute smallest value of F. Minimization algorithms however are fundamentally related to the number and distribution of *local* minima. In the best case, there would be only one local minimum, and only a relatively small number of evaluations of F would be needed to find the global minimum. The actual number of evaluations would depend rather strongly on the number of dimensions and on the shape of the function, but very roughly only about 10^3 evaluations would be needed. In the worst case the local minima would be distributed randomly, and a brute-force search would be necessary to find the global minimum. The number of evaluations needed would then be $(N_{search})^{N_{dim}}$, where N_{search} is the number of search points for each dimension, and N_{dim} is the number of dimensions. With typical values of 40 for the number of dimensions and 200 for the number of search points, it is evident that the worst-case problem is unsolvable.

B Optimization in atomic-scale modeling

In atomic-scale modeling, optimization problems are particularly important. The energy E of a system of atoms is a function of the geometric coordinates X_i of the atoms, and the equilibrium geometry is the set of coordinates with the minimum energy. This is a multi-dimensional minimization problem. Energy minimization is perhaps the most widely used numerical problem in atomic-scale modeling. Next, in

many quantum mechanics models, the quantities of interest depend on a set of electron numbers N_i, but the electron numbers appear in an equation which can not be solved explicitly for N_i. This is the widely-encountered self-consistency problem, and this is a multi-dimensional root-finding problem. Next, in many atomic-scale modeling problems one has a set of free parameters or empirical parameters S_i, and one wants to find the set of parameters that give the "best" calculated properties P_k. The concept of the "best" properties is usually quantified as the minimum value of a least-squares sum, and so this is a least-squares problem.

In many energy minimization problems, one is interested in how the energy depends on a single geometric coordinate. For example, for bulk or crystalline systems one is often interested in the energy E as a function of the atomic volume V. This is a one-dimensional minimization problem. The very important Fermi energy E_F, which determines the number of electrons occupying each energy eigenvalue, is defined by an equation which can not be solved explicitly. This one-dimensional root-finding problem is particularly difficult because the Fermi function is extremely nonlinear. Finally, the numerical calculation of elastic coefficients, and the very closely related vibration frequencies, depend on the properties of the energy as a function of a specific geometric coordinate or mode of vibration. This one-dimensional numerical derivative problem, although not strictly an optimization problem, uses many of the same concepts as other optimization problems.

We should also mention the two very important research areas of molecular dynamics and Monte Carlo simulations. Our atomic-scale modeling program does not include these two types of simulations. However, in the context of optimization problems, both molecular dynamics and Monte Carlo simulations can be interpreted as generalizations of the energy minimization problem. In addition to using the energy $E = E(X_i)$ to find the equilibrium geometry, molecular dynamics and Monte Carlo simulations are used to calculate a wide variety of kinetic and thermal properties, each of which can be calculated from some property of the function

$E(X_i)$. A discussion of these techniques is outside the scope of this report. However, some of the results in this chapter might be useful in these areas.

C Logistical issues

Numerical optimization problems have a relatively long history and are well-understood (see Ref. [17], [25]). The local minimization problem is considered to be a "closed case", with *variable metric* algorithms often providing the best performance. The local root-finding problem is also considered to be a closed case, with variants of either the *Newton* algorithm or the *Broyden* algorithm often providing the best performance. The self-consistency problem is still widely discussed in the literature, but this is usually in the context of making relatively small modifications to improve the performance. The local least-squares problem is more complicated, but the *Levenberg-Marquardt* algorithm works so well that for most practical purposes this is also a closed case.

The global problems are somewhat different. The global root-finding problem does not appear to be of major importance; the self-consistency problem is almost always treated as a local problem, as the existence of multiple solutions is not physically reasonable. The global minimization problem is very important in many different fields. Although still an active area of discussion in the literature, it is now fairly well-understood, with several types of algorithms available. The global least-squares problem is *not* well-understood and it is not widely discussed in the literature. Our need for an efficient global least-squares algorithm has led us to develop a global modification of the Levenberg-Marquardt algorithm.

The one-dimensional problems for both minimization and root finding are also a closed case. In this area there are many algorithms available, and the selection of the algorithm usually depends on the specific problem. In practice reliability issues, such as handling unexpected conditions, are often as important as performance issues. In fact, if performance is critical for a one-dimensional problem,

the strategy is usually to use the previously evaluated points to construct a curve such as a spline, and in this case the problem becomes less of a minimization problem and more of a spline interpolation problem. Global problems in one dimension are also not problematic, as it is usually possible to solve such problems using a brute-force search.

D Least-squares residual

The physical properties calculated by empirical models depend on a set of parameters S_i. The objective of empirical modeling is to find the best set of parameters S_i. This is done by selecting a set of physical properties P_k, called *fitting properties*, and constructing a least-squares sum:

$$R(S_i) \overset{\text{first attempt}}{=} \sum_k P^k_{weight} \cdot \left(P^k_{calc}(S_i) - P^k_{ref}\right)^2 \tag{28}$$

For each fitting property, there is a weight factor P_{weight} which represents the relative importance of the property, the value P_{calc} calculated by the model, and a reference value P_{ref} which is the desired value of the property. Ideally, the reference values would be the experimental values of the physical properties. However, due to the need for a large number of fitting properties, and the need for a uniform technique to be used for the reference values, the reference values are usually obtained from first-principles or ab-initio calculations such as density functional theory.

The scalar value R is called the *residual*, as it represents the amount by which the calculated values differ from the reference values, and of course if all the calculated values are equal to the reference values then the residual is zero. The form in eq. 28 is widely used, and it is at least sufficient for a least-squares optimization. However, in this form the actual numerical value of the residual is rather meaningless, particularly if properties with different physical units are used.

Our final form for the residual R is:

$$R(S_i) \overset{\text{final form}}{=} 1000 \cdot \sqrt{\frac{1}{N_P} \cdot \sum_k \left(P^k_{weight} \cdot \frac{P^k_{calc}(S_i) - P^k_{ref}}{P^k_{scale}} \right)^2} \tag{29}$$

For each property, we now also have a characteristic scale P_{scale}. The values P_{calc}, P_{ref}, and P_{scale} should all have the same physical units, and then the weight P_{weight} is unitless, and also the residual R is unitless. The weight factors are now squared, which results in a more intuitive interpretation of weight as a relative importance. Dividing by the number of properties N_P prevents the numerical value of the residual from scaling with the number of properties. That is, if we double the number of terms in the sum in eq. 28, this has the undesirable effect of doubling the value of the residual. The square root allows for the residual to scale linearly with the difference $P_{calc} - P_{ref}$.

With these modifications, the residual is now a type of average of the differences $P_{calc} - P_{ref}$. This means that the actual numerical value of the residual is now physically meaningful as the average deviation of the calculated values from the reference values. The factor of 1000 in eq. 29 is used to obtain a "user-friendly" numerical value. This factor is chosen so that a residual of 1.0 corresponds to an accuracy of 1 part in 1000; this is approximately the level called "chemical accuracy", which is something of a reference accuracy for atomic-scale calculations. Our interpretation of various values of the residual, which has been quite useful in practice, is shown in Figure 6.

The use of a residual with a physically meaningful value is of critical importance. For example, suppose that there are two research groups that are competing to develop the best numerical model for some physical system. How can we decide which group has the best model? With the residual in eq. 28 there is no way to decide unless each group uses *exactly* the same set of fitting properties P_k with *exactly* the same reference values P_{ref}. With a physically meaningful residual as in eq. 29 we can make such decisions, even if different fitting properties and

residual value	interpretation
≈ 1000	The calculated values are within an order of magnitude of the reference values. Values around 1000 are often encountered during optimization, since the dependence of the residual on the parameters is extremely nonlinear.
≈ 100	The calculated values are accurate to about 1 part in 10. This is a good value for an initial or starting set of parameters, before optimization. If the initial residual is much larger, it is likely that the optimization will fail to find the minimum in a reasonable amount of time.
≈ 50	We have chosen this value somewhat arbitrarily as meaning that the physical model is reasonable, or similarly that the values of the parameters are reasonable. After much experimentation this still seems to be a good threshold value.
≈ 20	This about the best value that we have obtained for a full optimization. At this level, not only are the individual calculated values accurate, but also comparative values such as energy differences and trends and patterns are also accurate. This is very desirable because it suggests that the model will be accurate for molecular dynamics and thermodynamics simulations, which depend on energy differences.
≈ 10	This is approximately the accuracy of the first-principles calculations used to obtain the reference values. Any attempt to obtain a residual less than the accuracy of the reference values is misguided since one will be no longer be exploring real-world values, but rather the limitations of the first-principles calculations.
≈ 1	The calculated values are accurate to about 1 part in 1000. Some individual properties might have a meaningful residual at this level if their reference values are sufficiently accurate.

Figure 6. Interpretation of various values of the least-squares residual R.

different reference values are used. There are a variety of similar situations where one needs such a comparison or competition. For example, one might want to apply the same model to several different systems and then ask which system the model works best for. Or, one might want to test a new modification to an existing model to see if the modification provides any improvement. Finally, because the residual R is a type of average, one can identify a residual R_k for each property. This can be used to identify how well a model works for an individual property.

We should also mention that one might want to modify the residual to account for the number of empirical parameters used in the model. Formally, if the region near the minimum is linearized this gives a set of linear equations:

$$P^k_{calc}(S_i) = P^k_{ref}$$

which can be solved exactly for any number of properties up to the total number of parameters N_S. In practice there are nonlinear effects, but there is still the informal sense that if one has, for example, 200 properties and 40 parameters, that only 160 properties have been fit in a non-trivial way. This suggests the use of the following modification in eq. 29:

$$\sqrt{\frac{1}{N_P}} \rightarrow \frac{N_P}{N_P - N_S} \cdot \sqrt{\frac{1}{N_P}}$$

This will increase the value of the residual R, indicating a poorer fit. Loosely speaking, this modification penalizes a model for using too many empirical parameters. Note that for $N_S \geq N_P$ the residual is meaningless, as it should be, indicating that nothing has been fit in a non-trivial way.

E Global least-squares

For most of the optimization problems in our program, efficient algorithms are already available, as discussed in Section C. This is not the case for the global least-squares problem. There are in general two approaches to the global problem. The first is to treat the residual R as a scalar value, and to use a global

minimization algorithm. This has the advantage that one can use the pre-existing algorithms without the need to develop a new algorithm. However, there are N_p terms in the least-squares summation, and if the summation is performed explicitly then a large amount of information about the behavior of the individual terms is lost. So, the second approach is to adapt the Levenberg-Marquardt algorithm to the global problem. Other atomic-scale modeling problems similar to our own seem to prefer the first approach of treating the global problem as a scalar problem. However, we have found that the second approach of treating the global problem as a least-squares problem gives a faster optimization. In this section we discuss our global modifications to the Levenberg-Marquardt algorithm.

The most obvious starting point for the global modification is a to send successive sets of parameters S_i to the local L-M algorithm. It is useful in this context to use the terminology "base camp" and "search team" to refer to the relevant sets of parameters. The base camp S_{base} is the initial or starting set of parameters. From this base camp we send out a search team S_{search}, which is a new set of parameters. The search team is used by the L-M algorithm, which moves S_{search} downhill to a local minimum. Now, if the local minimum is acceptable, then we move the base camp to the local minimum. If the local minimum is not acceptable, then we simply ignore it and send out a new search team. For this there are two essential procedures that must be specified: first, how to construct a new set S_{search} from a current set S_{base}, and second, how to decide if a local minimum is acceptable. For the first procedure we define a scalar S which can be interpreted as the distance between S_{search} and S_{base}:

$$S = \sqrt{\frac{1}{N_S} \cdot \sum_k \left(\frac{S^k_{search} - S^k_{base}}{S^k_{scale}} \right)^2} \tag{30}$$

If we assume that the global minimum is more likely to be a small distance from S_{base} than a large distance from S_{base}, then it is reasonable to choose successive sets S_{search} using a random distribution where small values of S are more likely and

large values of S are less likely.

In the absence of any directional information about the distribution of local minima, the direction from S_{base} to S_{search} should be chosen randomly. Our algorithm for the construction of S_{search} from S_{base} is then:

$$S^i_{search} = S^i_{base} + S \cdot \left(\frac{\sum_k r^k_{-1,+1} \cdot r^k_{-1,+1}}{N_S} \right)^{-\frac{1}{2}} \cdot r^i_{-1,+1} \cdot S^i_{scale} \tag{31}$$

with:

$$S = -S_{global} \cdot ln\,(r_{0,1}) \tag{32}$$

where $r_{a,b}$ is a random number with uniform distribution over the interval from a to b. The factor $r^i_{-1,+1}$ provides the random direction, and the square root factor acts as a normalization coefficient to satisfy the constraint in eq. 30. Our specific choice for the random value of S in eq. 32 is the exponential distribution [17], where S_{global} is a unitless "half-life" constant which represents the expected range over which the local minima are distributed. For example, a value of $S_{global} = 0.30$ indicates that the search parameters will differ from the base parameters by about 30%. If one has some further knowledge about the distribution of the local minima, then it would of course be reasonable to use a different random distribution in place of eq. 32. Finally, although it is evident from the context of the discussion, we should emphasize that the array r^i or r^k refers to the same array of random numbers for each of its 3 appearances in eq. 31..

For the second procedure of deciding whether a local minimum is acceptable, we simply use the rule that it is acceptable if it is the best local minimum found so far. An alternative would be to move the base camp to the search team using a random probability that depends on the value of the residual at the local minimum. This would take us into the area of thermal techniques such as simulated annealing, which interpret the residual as a type of physical energy barrier. Perhaps the most important feature of our algorithm is that it does *not* use thermal techniques. The dependence of the residual R on the parameters S_i is highly nonlinear, even in

regions where the values of the parameters are reasonable. This suggests that thermal techniques are not appropriate for these types of problems, because the residual barriers are too large. If one wishes to adapt this algorithm to problems where a thermal interpretation is more appropriate, one can use a probability for moving to the local minimum that depends on the difference between two appropriate values of R. Indeed, we have used this thermal adaptation at times, and although it certainly adds flair to the algorithm, it does not appear to be useful for our particular problem.

Finally, we should point out that this algorithm can be easily and highly parallelized. For these types of atomic-scale modeling problems, the derivatives of the least-squares terms can not be evaluated analytically. This means that some type of forward difference must be used to calculate the Jacobian matrix elements $J_{ki} = \frac{\partial P_k}{\partial S_i}$, which are used by the Levenberg-Marquardt algorithm. There will be a forward difference for each of the parameters S_i, and since each forward difference is independent of the others, we can parallelize the calculation of the Jacobian by sending one forward difference to each processor. Coincidentally, the number of parameters N_S is just about same as the number of processors available on a modern multi-processor computer. One could also parallelize the algorithm by sending an entire local optimization to each processor, with some minor modifications to account for the fact that each of these local optimizations must run independently of the others in order to be parallelized.

F Distribution of local minima

As with any global algorithm, it is the validity of the assumptions about the distribution of local minima, and not the creativity of some anthropomorphic analogy, that determines whether the algorithm is useful for a particular problem. Indeed, it is a fair criticism of some global algorithms that too much emphasis is placed on such anthropomorphic analogies. In this section we briefly discuss the

justification of our assumptions about the distribution of local minima.

For atomic-scale modeling problems, the empirical parameters are usually chosen to have as simple a physical interpretation as possible. For example, an empirical parameter might represent the spatial extent of the distribution of electrons around a Silicon atom, which is known to be about 5Å. It is then expected that the optimized set of parameters is more likely to have a value in the range 4Å - 6Å, less likely to have a value in the range 3Å - 7Å, and unlikely to have a value outside this range. This suggests that the exponential distribution in eq. 32 is valid.

It is not enough just to *have* such a distribution of local minima. One must also *start* somewhere inside this distribution. That is, in eq. 31 the construction of the search parameters S_{search} from the base parameters S_{base} implies that the starting parameters are relatively close to the global minimum. This is consistent with our use of empirical parameters which have a simple physical interpretation; accurate starting values for such parameters can usually be obtained.

We should also mention the important role of the characteristic scales S_{scale} in eq. 30. The purpose of the scales is to be able to construct a scalar value for the distance between two sets of parameters. Unfortunately, unlike in real three-dimensional physical space where distance is well-defined, there is no such well-defined distance for a set of parameters S_i. It is the scales that define the concept of distance, or more formally the *metric*, for the parameters.

Next, as we have discussed previously, based on our own observations, the dependence of the residual on the parameters is extremely nonlinear. Continuing our example, there might be a small (good) residual at 5.5Å, but a very large (bad) residual at 4.5Å. This suggests that the numerical values of the residual do not contain any useful information about the distribution of local minima. This is consistent with our algorithm in eq. 31: new parameters S are constructed only from other parameters S and not from any residual values R.

It should be possible to develop better algorithms for the global least-squares

problem. In one of the very few articles on this subject, Velázquez et. al. [23] have suggested that, for a large class of problems, the numerical values of the residual contain information about the distribution of local minima. Their technique, called selective minimization, is based on the observation that "smallest residual" or "smallest deviation" or "smallest error" problems are a special type of least-squares problems, distinguished from the general least-squares problem by the fact that the global minimum has a very small residual. It is evident that empirical parameter modeling is just such a smallest-residual problem.

G Gaussian fill

Global optimization algorithms have a tendency to return to the same local minimum over and over again. For a molecular dynamics or Monte Carlo simulation this might be a good thing, because physical properties such as vibration frequencies and transition rates can be calculated from the probability of returning to a local minimum. For empirical modeling this is not a good thing, because we are interested only in finding the global minimum as quickly as possible. This can be stated more formally by saying that for empirical modeling there is no physical significance to the dynamical path taken by the algorithm. Returning to the same local minimum is simply a waste of time.

A simple and effective solution to this problem has been developed recently by Parrinello et. al. [24]. Their solution is to add to the residual R a relatively narrow Gaussian function centered at each of the previously-found local minima L:

$$R_{fill}(S_i) = R(S_i) + \sum_L G(S_i, S_{L,i}) \tag{33}$$

The entire summation is zero except when the current set of parameters S_i is very close to one of the local minima $S_{L,i}$. The Gaussian functions act to fill up each local minimum; when a minimum is sufficiently filled it is no longer a minimum, and the optimization algorithm will no longer return to it. In theory, filling the local

minima is problematic because the Gaussian terms are history-dependent. That is, for a specific set of parameters S_i, we can have $R_{fill} = R$ early in the optimization, and $R_{fill} \neq R$ later in the optimization. In practice this is not a problem: R_{fill} is never interpreted as the official value of the residual; it is only a raw value used by the optimization algorithm.

Our specific form for the Gaussian functions in eq. 33 is:

$$G(S_i, S_{L,i}) = R(S_i) \cdot r_{fill} \cdot \exp\left(-s_{fill} \cdot S(S_i, S_{L,i}) \cdot S(S_i, S_{L,i})\right) \tag{34}$$

with:

$$S(S_i, S_{L,i}) = \sqrt{\frac{1}{N_S} \cdot \sum_k \left(\frac{S_k - S_{L,k}}{S^k_{scale}}\right)^2}$$

Here $S(S_i, S_{L,i})$ is just a scalar value for the distance between the current set of parameters S_i and the local minimum $S_{L,i}$. This form introduces two new unitless constants s_{fill} and r_{fill} for the width and height of the Gaussian functions. The value of s_{fill} should be close to (or less than) the expected separation between local minima, so that the Gaussians from different local minima do not overlap. We use a value of $s_{fill} = 0.02$, but of course this should not be taken to be a "universal" value. The value of r_{fill} should be close to (or less than) the expected depth of the local minima; we use a value of $r_{fill} = 0.20$. Finally, we have included the residual R as a factor for the Gaussian functions in eq. 34; this seems to be necessary in order to interpret r_{fill} as a fixed constant.

H Success-failure algorithms

We have discussed previously that if one has a one-dimensional problem, it is usually not suitable to use a multi-dimensional algorithm with the number of dimensions set to one. In this section we discuss the one-dimensional algorithm that we use for minimization. The algorithm is due to Rosenbrock [26]. The input consists of X_{init}, which is the initial or expected X-value for the minimum, and

X_{change}, which is the initial or expected change in X_{init}. The algorithm also monitors the variables X_{best} and F_{best}, which correspond to the best or smallest evaluated value of $F(X)$. To get things started, one evaluates $F(X)$ at the points X_{init} and $X_{init} + X_{change}$. Each evaluation of $F(X)$ is then called a success if $F < F_{best}$ and a failure if $F \geq F_{best}$. After the evaluations at X_{init} and $X_{init} + X_{change}$, one sets the variable step size $X_{step} = X_{change}$, and then constructs a trial value of X in one of two ways, depending on whether the most recent evaluation of $F(X)$ is a success or a failure:

$$\begin{aligned} &\text{success:} \quad X_{trial} = X_{best} + c_{expand} \cdot X_{step} \\ &\text{failure:} \quad X_{trial} = X_{best} - c_{contract} \cdot X_{step} \end{aligned} \tag{35}$$

The coefficients c_{expand} and $c_{contract}$ are expansion and contraction coefficients. They are constrained by the conditions:

$$\begin{aligned} c_{expand} &> 1.0 \\ 0.0 < c_{contract} &< 1.0 \end{aligned} \tag{36}$$

After X_{trial} is assigned, the value of X_{step} is updated to the new step size $X_{step} = X_{trial} - X_{best}$. The function $F(X)$ is then evaluated at the trial point X_{trial}, the variables X_{best} and F_{best} are updated, and the entire process is repeated.

The success-failure algorithm is included in our discussion of optimization techniques because of the serious errors that can result from the use of a one-dimensional minimization or root finding algorithm. First, many one-dimensional algorithms require the specification of a range of X-values in which the minimum or root is located (see Ref. [17]). Based on our own experience, we feel that the use of any specified-range algorithm is unacceptable for the physical models discussed in this report. The problem with such algorithms is that they can return the upper or lower bound of the range as the minimum. For example, if we attempt to minimize the function $F = (X - 4)^2$ using the range $X = [6, 20]$, we might be told that the function has a minimum at $X = 6$. It would of course be

possible to add a separate algorithm to search for a range that is guaranteed to contain a local minimum, or to modify the algorithm to report an error message if the lower or upper bound is returned as the minimum. In practice these modifications add an unnecessary level of complexity to what should be a simple problem. The success-failure algorithm uses only an initial point X_{init} and an initial step size X_{change}; it does not require a specified range for the minimum.

Next, many one-dimensional algorithms use polynomial interpolation to reduce the number of function evaluations needed to find the minimum. Unfortunately, this introduces a large number of unexpected conditions that must be accounted for. These include a polynomial with a maximum rather than a minimum, a polynomial with a minimum outside the range of X-values used to construct the polynomial, and polynomial that is a straight line. Also, if such an algorithm is very close to the minimum, the polynomial is very close to a straight line, and division by zero can cause the algorithm to fail. This requires additional modifications to account for the final stage of the minimization. The success-failure algorithm updates X_{best} using only the expansion or contraction step in eq. 35. There are no such unexpected conditions associated with the update of X_{best}, and no such modifications for the final stage of the minimization.

This does not mean that the success-failure algorithm can not be modified to reduce the number of function evaluations needed to find the minimum. It means that such modifications are much less likely to cause errors than they would be if made to a different algorithm. This is because the success-failure algorithm can serve as a framework for a polynomial interpretation algorithm. The expansion and contraction steps in eq. 35 will converge to the minimum as long as the conditions in eq. 36 are satisfied, even if the expansion and contraction factors take on different values during the minimization. The strategy here is to use polynomial interpolation to *suggest* or *recommend* a step size to be used in eq. 35, and then use the success-failure framework to decide whether to accept this step, or to reject it

and revert to a default expansion or contraction factor.

I Fermi energy algorithms

Our final optimization technique is the atomic-scale modeling problem of the calculation of the occupation numbers from the energy eigenvalues. The input to the problem consists of an array $\{E_i\}$ for the energy eigenvalues of an atomic-scale system. Sizes in the 10000s of eigenvalues are typical. The output consists of the array $\{N_i\}$ for the number of electrons occupying each eigenvalue or eigenstate. The occupation numbers N_i are specified by the equation:

$$N_i(E_i) = n_{elec} \cdot \exp\left(-\frac{(E_i - E_F)}{E_T}\right) \cdot \left(1 - \exp\left(-\frac{(E_i - E_F)}{E_T}\right)\right)^{-1} \qquad (37)$$

where n_{elec} is a fixed parameter (input) for the maximum number of electrons allowed to occupy a single state, and E_T is a fixed parameter (input) for the "thermal energy" of the electrons. The Pauli exclusion principle requires that $n_{elec} = 1$ or $n_{elec} = 2$ depending on whether the physical model treats electron spin explicitly. The value of E_T is typically very roughly on the order of 1 part in 10^6, assuming that a characteristic scale for the eigenvalues is available. We should point out that E_T does not represent the actual real physical temperature of the system of atoms. The physical temperature is usually associated with the motion of the nuclei of the atoms, as in a molecular dynamics or Monte Carlo simulation.

The remaining variable in eq. 37 is the scalar Fermi energy E_F. The value of E_F is specified by the constraint:

$$N_{elec} = \sum_i W_i \cdot N_i(E_F)$$

where N_{elec} (input) is the total number of electrons in the system. The weight factors W_i are all $W_i = 1$ for a system without periodic boundary conditions. However, for a periodic system such as a crystal or surface it is necessary to treat the general case of arbitrary weight factors. The arrays $\{E_i\}$, $\{N_i\}$, and $\{W_i\}$ are

all multi-dimensional. However, the scalars E_F and N_{elec} are both one-dimensional. From the perspective of the numerical algorithm, the energies E_i and weights W_i are treated as fixed input, and the scalar E_F is treated as an unknown. The relevant equation for the algorithm is then $\sum_i W_i \cdot N_i(E_F) - N_{elec} = 0$, which is a one-dimensional root-finding problem.

Before proceeding, let us clarify the role of this problem in atomic-scale modeling with an informal example. Consider the set of eigenvalues $\{-12.000, -10.000, -8.000, -6.000\}$ for a system with a total of 4 electrons. Loosely speaking, as a first attempt we want to put two electrons into each of these eigenvalues or eigenstates, giving the array of occupation numbers $N_i = \{2.000, 2.000, 0.000, 0.000\}$. In this simple example we can assign N_i without actually calculating E_F. However, if we gradually increase the value of -10.000 and decrease the value of -8.000, this first attempt at assigning N_i will result in values that do not have a smooth dependence on E_i. The distribution in eq. 37 is introduced to restore this smooth dependence. In our example this second attempt might result in the occupation numbers $N_i = \{1.999, 1.999, 0.001, 0001\}$. That is, most of the occupation numbers will either be very close to zero or very close to n_{elec}, with the possibility of having some intermediate values if some of the energies E_i are very close to each other.

The Fermi energy problem is included in our discussion of optimization techniques because it is especially prone to errors or bugs. The root-finding function $\sum_i W_i \cdot N_i(E_F) - N_{elec}$ is extremely flat in regions where E_F is not close to one of the energies E_i. In fact, because of the limitations of floating-point storage, the function is *exactly* flat in these regions. In practice a root-finding algorithm will usually fail in these regions; a perfectly flat region contains no information about how to proceed toward a root. This can be developed more formally by determining the range over which the function is *not* flat. For this we need the *machine accuracy* ε, which is usually defined as the smallest number for which 1 and $1 + \varepsilon$ can be

distinguished from each other. For our purposes this means that the root-finding function is non-flat for $N_i > \varepsilon$ and $N_i < n_{elec} \cdot (1 - \varepsilon)$. Using these values in eq. 37 shows that the function is non-flat in the regions:

$$E_F \approx E_i \pm \ln(\varepsilon^{-1}) \cdot E_T \tag{38}$$

This important equation shows that $\ln(\varepsilon^{-1})$ is not large enough to extend the non-flat regions from one value of E_i to another. With 64-bit floating point storage, the logarithm in eq. 38 has taken the range of non-flat coverage from a factor of $\varepsilon^{-1} \approx 1 \cdot 10^{16}$ to a factor of only $\ln(\varepsilon^{-1}) \approx 40$. Since E_T is required to be small, and since typical separations between energy eigenvalues are on the order or 1 part in 10^1, we have shown that exactly flat regions, which are expected to cause a root-finding algorithm to fail, are common for the Fermi energy problem.

We have experimented with several possible modifications to prevent a root-finding algorithm from failing. Our first attempts were to use modified root-finding algorithms that could handle exactly flat regions. Our next attempts were to identify cases where we could assign the occupation numbers without actually calculating E_F. This will work for any physical system that has a well-defined band gap, i.e. a band gap larger than $\ln(\varepsilon^{-1}) \cdot E_T$. However, we found that these attempts are prone to errors or bugs, especially for periodic systems. Our solution is to return to an *unmodified* root-finding algorithm. The trick is to very carefully assign an initial value for E_F, so that the root-finding algorithm always starts in a non-flat region and never has a chance to enter a failure-prone flat region. We use a first pass through the array $\{E_i\}$ to find the elements of the array $E_{i_{lower}}$ and $E_{i_{upper}}$ that are the upper and lower bounds for the Fermi energy. Note that because of the need to treat the weight factors W_i for periodic systems, we can not use a trivial assignment such as $i_{lower} = \frac{N_{elec}}{n_{elec}}$. These elements $E_{i_{lower}}$ and $E_{i_{upper}}$ can be identified by the condition:

$$\sum_{i=1}^{i=i_{lower}} W_i \cdot n_{elec} > N_{elec} - \varepsilon_{safe} \cdot E_{scale}$$
$$\sum_{i=1}^{i=i_{upper}} W_i \cdot n_{elec} > N_{elec} + \varepsilon_{safe} \cdot E_{scale}$$

where ε_{safe} is a small tolerance with $\varepsilon_{safe} \gg \varepsilon$, and E_{scale} is the characteristic scale of the energy eigenvalues. The initial value of E_F for the root finding algorithm is then just $E_F = \frac{1}{2}\left(E_{i_{lower}} + E_{i_{upper}}\right)$. We have found that this modification is very stable for both non-periodic and periodic systems over a large range of system sizes.

CHAPTER VI

ENVIRONMENT-DEPENDENT TECHNIQUES

A Introduction

In Chapter II we discussed what can be called a "standard model" for a self-contained algorithm to calculate the energy of an arbitrary configuration of atoms using only two-center integrals, or more generally parameterized functions which represent two-center integrals. The previous discussion actually already introduced several of the environment-dependent concepts that are the subject of this chapter. In some ways, we came close in Chapter II to spelling out an environment-dependent model.

The two key words or phrases associated with our model are "environment-dependent" and "self-consistent". These are not just buzzwords; but are important to describe the manner in which our model compares to other competing models. Environment-dependent refers in general to any interaction beyond those of a two-center model. In our model, the environment-dependent effects account for both the three- and four-center integrals that are not treated in a two-center model. While there are a few competing models that include environment-dependent effects, it is the full iterative self-consistent treatment of charge redistribution effects that set our model apart from competing models.

Our discussion in this chapter is out of necessity less refined than our discussion of two-center techniques in Chapter II. There, we were able to provide "line-by-line" derivations, and we were also able to show "term-by-term" correspondence of the components of our model with the components of

first-principles models. Perhaps the most important point in this respect is that the environment-dependent part of our model is more phenomenological, and that there is less opportunity here for such line-by-line and term-by-term derivations. This is due at least in part to the leading-edge nature of this research.

It is also very important for the purposes of this report as a dissertation to point out that the individuals involved in this research specialized in different areas of the project. My own work was more highly specialized in the *implementation* of the algorithms, the development of a *first-principles database,* and the *preliminary fitting* of the empirical parameters for C, Si, and Ge. A colleague, Dr. Ming Yu, specialized more highly in the subsequent fitting of C and Si, and the applications of the model to C and Si systems. As a result my discussion in this chapter is more oriented toward those areas in which I was more heavily involved.

B First-principles approach

Our discussion in Chapter II exhausted the types of mathematical objects that can be obtained from the bundle of approximations that comprises what can alternately be called "tight binding" or "two-center" theory. If we were to ask hypothetically what the most evident extensions or modifications to this theory would be, from the perspective of a two-center model only, there are two apparent directions that we could take. The first would be to modify the pairwise repulsive energy to include higher-order terms, the most likely of which would involve the bond angles θ_{ijk} associated with each triplet of atomic nuclei. Recall that the "derivation" of the repulsive energy is on very weak ground, as it represents a composite term which is known to be very complicated in first-principles treatments:

$$E_{repulsive} = E_{nuclei\text{-}nuclei} - E_{electrons\text{-}electrons}$$

The second direction would be to treat the three-center integrals; along with the repulsive energy, these integrals are really the only mathematical objects that we

are free to work within the central field approximation.

There is however an increasing consensus that this hypothetical approach is not productive. Bond-angle terms and other classical modifications to the repulsive energy would result in a sharp increase in the number of parameters, while at the same time the classical nature of such modifications would work against the concept of having an electronic structure model. Three-center integrals of course can not be criticized as being classical in nature; however, they also would suffer an unacceptable increase in the number of parameters. Furthermore, the growing consensus is that these integrals are simply not the "weakest link" in two-center models.

The consensus from both the theoretical and practical approaches is that the *charge redistribution*, or more generally some modification involving the charge density, is the most important item in the development of models that approach the accuracy of first-principles calculations, while at the same time maintaining the fast speed that allows one to study larger systems. This concept of charge redistribution is of course not present in two-center models, having been lost in the various layers of approximation; the matter must be approached from a first-principles perspective. Our discussion follows closely that of our own recent publication [30]. We begin with the many-body Hamiltonian [29]:

$$H = -\sum_{l} \frac{\hbar^2}{2m} \cdot \nabla_l^2 + \sum_{il} v(\mathbf{r}_l - \mathbf{R}_i) + \sum_{ll'} \frac{e^2}{4\pi\varepsilon_0 \cdot r_{ll'}} + \sum_{ij} \frac{Z_i \cdot Z_j \cdot e^2}{4\pi\varepsilon_0 \cdot R_{ij}}$$

where l and l' index the electrons, and i and j index the nuclei. Z refers to the number of electrons associated with the neutral atom; for the purposes of our empirical model, which uses a valence approximation, Z will refer to the number of valence electrons.

C On-site terms

When the above Hamiltonian is treated in a one-particle approximation, one obtains an expression for the on-site terms, which serves as a starting point for our environment-dependent model:

$$H_{i\alpha,i\alpha} = \varepsilon^0_{i\alpha} + u_{i\alpha} + u'_{i\alpha} + v_{i\alpha} \tag{39}$$

The individual terms in this expression refer to various interactions involving the orbital indexed by α and associated with the atom indexed by i. $\varepsilon^0_{i\alpha}$ refers the interaction with its own nucleus, $u_{i\alpha}$ the interactions with orbitals at its own site, $u'_{i\alpha}$ the interactions with orbitals at other sites, and $v_{i\alpha}$ the interactions with nuclei at other sites; $\varepsilon^0_{i\alpha}$ also includes the kinetic energy. At this point there are several directions that one could proceed in, depending on the extent to which one wants to treat the self-consistency problem, which describes the charge redistribution.

In our model we choose a rather ambitious treatment requiring an iterative numerical treatment, i.e. a root-finding algorithm, which is the numerical or computational equivalent of the self-consistency problem. However, we avoid treating the charge density with a three-dimensional grid or mesh, which would slow the model down to an unacceptable level. Instead, we have chosen to treat the charge density using the electron numbers N_i associated with each site i. Our semi-empirical treatment of the terms in eq. 39 is:

$$\varepsilon^0_{i\alpha} = \varepsilon_{i\alpha} - Z_i \cdot U_i$$

$$u_{i\alpha} = N_i \cdot U_i$$

$$u'_{i\alpha} + v_{i\alpha} = \sum_{k \neq i} \left(N_k \cdot V_N(R_{ik}) - Z_k \cdot V_Z(R_{ik}) \right)$$

In these expressions, $\varepsilon_{i\alpha}$ is the traditional on-site energy corresponding to the eigenvalues of the isolated atom. If one wants to think of the model as a modification or extension of a traditional two-center model, then we can begin to

think of the Hamiltonian in the form:

$$H_{env} = H_{trad} + \text{modifications}$$

where *env* refers to our environment-dependent model, and *trad* refers to a traditional "tight-binding" model.

Mathematically, the scalar values U, and the functions $V_N(R)$ and $V_Z(R)$ can be discussed from different perspectives. One approach is to start with U (which describes same-site i-i interactions), and then to treat $V_N(R)$ and $V_Z(R)$ (which describe different-site i-k interactions) as generalizations of U. The other approach is start with $V_N(R)$ and $V_Z(R)$, and then to treat U as a special case of $V(R)$ for the same-site interactions.

In any event, the physical interpretation is that U describes the effective energy for electron-electron interactions at the same site, $V_N(R)$ describes the electron-electron interactions at different sites, and $V_Z(R)$ describes the orbital-ion interactions at different sites. In alternate treatments the scalar value U arises as part of the widely-used Hubbard model. Although our treatment of U still corresponds to a Hubbard model, in our model U is more of a starting point for the more important empirical functions $V_N(R)$ and $V_Z(R)$. For the computational problem $V_N(R)$ and $V_Z(R)$ are treated as parameterized functions, and U is treated as a parameter. Following our discussion in Chapter II, this parameterization is very important, as any modification of an existing model must not result in a sharp increase in the number of parameters. In Chapter II we saw that our two-center model uses roughly 20 parameters, representing nine parameterized functions. Our environment-dependent modification then adds two parameterized functions, resulting in a balanced increase in the number of parameters.

D Off-site terms

In our model the off-site terms are treated as generalizations of the on-site terms in Section C:

$$\begin{aligned} H_{i\alpha,j\beta}|_{i\neq j} &= \tfrac{1}{2}\cdot(\varepsilon_{i\alpha}+\varepsilon_{j\beta})\cdot K(R_{ij})\cdot S_{i\alpha,j\beta} \\ &+\tfrac{1}{2}\left((N_i-Z_i)\cdot U_i+(N_j-Z_j)\cdot U_j\right)\cdot S_{i\alpha,j\beta} \\ &+\tfrac{1}{2}\textstyle\sum_{k\neq i}\left(N_k\cdot V_N(R_{ik})-Z_k\cdot V_Z(R_{ik})\right)\cdot S_{i\alpha,j\beta} \\ &+\tfrac{1}{2}\textstyle\sum_{k\neq j}\left(N_k\cdot V_N(R_{jk})-Z_k\cdot V_Z(R_{jk})\right)\cdot S_{i\alpha,j\beta} \end{aligned} \tag{40}$$

The first property to note about this Hamiltonian is that the first line in eq. 40 is in the form of traditional two-center Hamiltonian:

$$H_{env} = H_{trad} + \text{modifications}$$

Here, H_{trad} is treated using a Hückel approximation, where each element of the Hamiltonian is constructed from its corresponding overlap element, as discussed in Chapter II:

$$H_{trad}|_{i\neq j} = \tfrac{1}{2}\cdot(\varepsilon'_{i\alpha}+\varepsilon'_{j\beta})\cdot K(R_{ij})\cdot S_{i\alpha,j\beta}$$

Following our discussion in Chapter II, although it is possible to interpret the Hückel approximation in terms of physical or theoretical arguments, one can also interpret this as a thoughtful set of constraints, which reduces the total number of parameters by re-using some of the overlap parameters for the Hamiltonian. In other words, one still has the *trad* part of the model in terms of the very general two-center functions $H_{ss\sigma}(R)$, $H_{sp\sigma}(R)$ etc.:

$$H_{ss\sigma}(R) = \varepsilon'_s\cdot S_{ss\sigma}(R)$$

The environment-dependent modifications to the off-site terms consist of contributions from the same-site i-i and j-j interactions (involving U) and from the different-site i-k and j-k interactions (involving V_N and V_Z). The rather unwieldy appearance of eq. 40 is a result of the requirement that H is symmetric or

Hermitian; eq. 40 is largely a straightforward symmetrization of the on-site formulas in Section C. The environment-dependent terms are also expressed in terms of the overlap elements $S_{i\alpha,j\beta}$, in the manner of a Hückel approximation. Again, at least symbolically, one can cast eq. 40 in a variety of interesting forms, such as:

$$H_{env} = \tfrac{1}{2} \cdot (\varepsilon_{i\alpha} + \varepsilon_{j\beta}) \cdot (K_{trad} + \text{modifications}) \cdot S_{trad}$$

which emphasizes the Hückel approximation, and:

$$(H_{ss\sigma}(R))_{env} = (H_{ss\sigma}(R))_{trad} + \text{modifications}$$

which emphasizes the two-center functions. Apart from suggesting that there are a variety of ways in which one could introduce more parameterized functions into the model, this line of symbolic analysis also has not resulted in any significant theoretical insight.

E Total energy

Following our discussion in Chapter II, in the formalism of first-principles models, the band energy contains an unavoidable double-counting of the energy between electrons and other electrons, resulting in an expression for the non-band contribution to the total energy as:

$$E_{non\text{-}band} = E_{nuclei\text{-}nuclei} - E_{double\text{-}count} \tag{41}$$

In two-center models, it is at this point that one introduces a pairwise repulsive energy to account for $E_{non\text{-}band}$. However, in our model we can explicitly evaluate the double-counting term:

$$E_{double\text{-}count} = \tfrac{1}{2} \textstyle\sum_i (N_i \cdot N_i - Z_i \cdot Z_i) \cdot U_i + \tfrac{1}{2} \sum_{ik\,:\,i \neq k} N_i \cdot N_k \cdot V_N(R_{ik})$$

Here we arrive at a very interesting feature of our model. While we could use a pairwise parameterized function for $E_{non\text{-}band}$, it turns out that we already have all

the ingredients in place to construct the total energy, without introducing any additional empirical parameters. This results from the explicit appearance of $V_N(R)$, which is known from theoretical considerations to contain a long-range R^{-1} term. When combined with $E_{n\text{-}n}$, which is of course also known to contain a long-range R^{-1} term, we can explicitly reproduce the cancellation of the long-range terms, and the resulting short-range "repulsive" energy.

Our expression for the total energy is then:

$$\begin{aligned} E_{tot} = E_{band} &+ \tfrac{1}{2}\textstyle\sum_{ik\,:\,i\neq k} Z_i \cdot Z_k \cdot V_C(R_{ik}) \\ &- \tfrac{1}{2}\textstyle\sum_i (N_i \cdot N_i - Z_i \cdot Z_i) \cdot U_i \\ &- \tfrac{1}{2}\textstyle\sum_{ik\,:\,i\neq k} N_i \cdot N_k \cdot V_N(R_{ik}) \end{aligned}$$

where the V_C term is equivalent to $E_{n\text{-}n}$, and the U and V_N terms are equivalent to $E_{d\text{-}c}$, in eq. 41. The potential $V_C(R)$ is just the Coulomb energy or potential:

$$V_C(R) = \frac{e^2}{4\pi\varepsilon_0 \cdot R}$$

This implies a requirement that $V_N(R)$ is equivalent to $V_C(R)$ at "large" distances, which in practice are any distances larger than the known short range over which the old repulsive energy acts:

$$V_N(R) \rightarrow V_C(R) \quad \text{for } R > R_{short}$$

The crux of the matter is that in our model, the long range terms do not *always* cancel; in fact, complete cancellation is a special case of the more general partial cancellation that occurs for systems with $N \neq Z$:

$$E_{tot} = \text{etc.} - \tfrac{1}{2}\textstyle\sum_{ik\,:\,i\neq k} \Delta N_i \cdot \Delta N_k \cdot V_C(R_{ik}) \quad \text{for } R > R_{short} \tag{42}$$

where $\Delta N = N - Z$. This of course is highly desirable, as long-range interactions are known to occur, and the inability to reproduce these interactions is a known limitation of two-center models.

F Functional forms

In Chapter II we argued that it is useful to separate the general concept of having a parameterized function from the specific parameterized functional form that is used. In principle we can think of searching for the best shape of the function as a whole, i.e. as in the manner of variational calculus. Nevertheless, due to the difficult global nature of the fitting problem, it is still necessary to specify a form with only a few parameters per function; high-order polynomials and other brute-force parameterizations are not acceptable.

As discussed in Section E, by requiring the same long-range R^{-1} behavior for both $V_C(R)$ and $V_N(R)$, one satisfies both the known theoretical properties of these functions, as well as the highly desirable "partial cancellation" of the electron numbers in eq. 42. If we return to our model for the Hamiltonian in eq. 40, we can see that this same long-range R^{-1} behavior is also implied for the function $V_Z(R)$:

$$V_Z(R) \rightarrow V_N(R) \rightarrow V_C(R) \quad \text{for } R > R_{short}$$

This results in a partial cancellation in the Hamiltonian elements as well as in the total energy:

$$H_{i\alpha,j\beta}|_{i\neq j} = \text{etc.} + \tfrac{1}{2}\textstyle\sum_{k\neq i} \Delta N_k \cdot V_C(R_{ik}) \cdot S_{i\alpha,j\beta} + \text{etc.} \quad \text{for } R > R_{short}$$

With these considerations in place, there are a only a limited number of ways that one can parameterize $V_N(R)$ and $V_Z(R)$.

In some of our earliest work on this model, we noted that for systems with no charge transfer ($N = Z$), which includes the stable crystalline structures of most elements, the Hamiltonian elements can be expressed as:

$$H_{i\alpha,j\beta}|_{i\neq j} = \text{etc.} + \tfrac{1}{2}\textstyle\sum_{k\neq i} Z_k \cdot \Delta V(R_{ik}) \cdot S_{i\alpha,j\beta} + \text{etc.} \tag{43}$$

where $\Delta V = V_N - V_Z$. We realized that by parameterizing the short-range function $\Delta V(R)$ directly, we could compare the new model to our extensive experience with

two-center models. That is, for $\Delta V(R) = 0$, the Hamiltonian reduces to a two-center model (for systems with no charge transfer); we can then use the analogy of a knob that can be used to "turn up" the magnitude of the environment-dependent modification, i.e. by turning up the magnitude of $\Delta V(R)$.

Of course, after becoming more familiar with environment-dependent models, one moves away from the need to always refer back to two-center models. It is important to point out however, that the identification of $\Delta V(R)$ was critical to our early understanding of the model. At that time, one of the chief criticisms of the model that we were using was that is was quite poor at reproducing the high-pressure phases of Si. There was a general consensus that this was due to the high coordination number; diamond Si has a small c.n. of 4, while the high-pressure phases (body and face-centered cubic) have coordination numbers of 8 and 12. Although these high-c.n. phases are not of material interest for Si itself, there was an increasing need to treat transition metals and other large-c.n. elements. Also, there is always the difficulty of treating C, which is known to be cause problems due to the very different chemistries of the c.n.=3 (graphite) and c.n.=4 (diamond) structures. In fact, some earlier two-center models attempted to remedy this situation by counting the coordination number of each atom, and using it to explicitly modify the total energy.

Following our identification of $\Delta V(R)$ in eq. 43, we realized that the summation over a short-range interaction has the effect of *counting the coordination number*, as:

$$H_{i\alpha,j\beta}|_{i\neq j} = \text{etc.} + N_{coord} \cdot \Delta V(R_{coord}) \cdot S_{i\alpha,j\beta} + \text{etc.}$$

where N_{coord} is some effective coordination number, and R_{coord} is some effective coordination distance. This early analysis suggested an important connection between the conventional wisdom of coordination-dependent effects, and the ability of our model to reproduce such effects without any artificial "bond-counting" functions.

In any event, returning to the actual functional forms for $V_N(R)$ and $V_Z(R)$, we settled on treating $V_Z(R)$ using a conventional polynomial $\times$ exponential, combined with a long range part:

$$V_Z(R) = \frac{e^2}{4\pi\varepsilon_0 \cdot R} \cdot (1 - (1 + B_Z \cdot R) \cdot \exp(-\alpha_Z \cdot R)) \tag{44}$$

Rather than treating $V_N(R)$ explicitly, we parameterize the short-range $\Delta V(R)$ using our customized "hyperbolic" functional form:

$$\Delta V(R) = (A_N + B_N \cdot R) \cdot \frac{1 + \exp(-\alpha_Z \cdot d_N)}{1 + \exp(-\alpha_Z \cdot (d_N - R))}$$

which of course results in $V_N(R)$ being well-defined as $V_N = \Delta V + V_Z$. It is readily seen that both $V_N(R)$ and $V_Z(R)$ have the appropriate long-range behavior. Finally, we also constrain the parameter A_N as:

$$A_N = U_i - (\alpha_Z - B_Z) \cdot E_0$$

which reproduces the appropriate limiting behavior $\lim_{R\to 0} V_N(R) = U$. Together with the use of the constant 1 instead of an additional parameter A_Z in eq. 44, this constraint is something of a "finishing touch" that is not of critical importance.

G Parameterization for Si

By performing an extensive parameter fitting, we have obtained a stable "official" set of parameters for Si. The details of the numerical optimization have been discussed in Chapter V. For this fitting we used a relatively large set of reference values, which were chosen with the goal of improving the transferability of the parameterization to a variety of large-scale systems. This includes cluster, bulk, and band structure properties which we will discuss in this and subsequent sections. For the clusters, we included the bond lengths and binding energies for 2-atom to 6-atom clusters. The comparison of the calculated and reference values for these clusters are shown in Table 1, which is taken with some minor modifications from our Ref. [30].

cluster	geometry	property	present work	ab-initio
Si_2	$D_{\infty h}$	bond length r (Å)	2.226	2.288
		binding energy e (eV)	−2.435	−2.499
Si_3	C_{2v}	r (Å)	2.284	2.357
		r (Å)	2.168	2.158
		e (eV)	−3.413	−3.574
	$D_{\infty h}$	r (Å)	2.141	2.167
		e (eV)	−3.427	−3.404
Si_4	D_{2h}	r (Å)	2.275	2.311
		e (eV)	−4.101	−4.242
	T_d	r (Å)	2.332	2.474
		e (eV)	−3.773	−3.659
	$D_{\infty h}$	r (Å)	2.116	2.156
		r (Å)	2.164	2.176
		e (eV)	−3.289	−3.367
Si_5	D_{3h}	r (Å)	2.207	2.306
		r (Å)	3.141	3.064
		e (eV)	−3.352	−4.452
	C_{4v}	r (Å)	2.209	2.275
		r (Å)	2.358	2.513
		e (eV)	−4.327	−4.266
	$D_{\infty h}$	r (Å)	2.082	2.133
		r (Å)	2.128	2.144
		e (eV)	−3.545	−3.534
	T_d	r (Å)	2.127	2.215
		r (Å)	3.475	3.617
		e (eV)	−3.334	−3.283
Si_6	D_{4h}	r (Å)	2.248	2.363
		r (Å)	2.639	2.734
		e (eV)	−4.698	−4.664
	D_{3d}	r (Å)	2.261	2.285
		r (Å)	2.948	3.208
		e (eV)	−3.896	−3.972
	D_{ih}	r (Å)	2.057	2.098
		r (Å)	2.072	2.134
		r (Å)	2.149	2.158
		e (eV)	−3.446	−3.464

TABLE 1

Results of our environment-dependent model for small Si clusters. The ab-initio values were calculated using the GAUSSIAN-98 software package, with the MPW1PW91 hybrid functional and the cc-pVTZ basis set.

One of the most important aspects of our choice of fitting properties is that we use not only the lowest-energy geometries for each cluster, but also several other geometries that do not have the lowest energy. For example, the lowest energy geometry of the Si_5 cluster is known to be the D_{3h} geometry. However, we also fit to the C_{4v}, D_{ih}, and T_d geometries. For each geometry, both for the reference values and for the calculated values, the geometry was fully relaxed, i.e. relaxed under the constraints of the required geometry of course.

The reasoning behind this strategy is quite important. One anticipates a model that can be used to study the statistical and thermodynamic properties of material (in our own work this usually takes the form of molecular dynamics calculations, but one could also anticipate the use of Monte Carlo methods). If one fits only to the lowest-energy geometries, it is likely that the resulting model will be less accurate for the calculation of items such as transition rates, etc. that involve non-equilibrium geometries. Our choice of geometries is designed to force the parameterization to address such materials. In fact, several of the geometries included in our fitting are not true local minima, having imaginary frequencies that lead to other geometries. I was motivated in this choice by the pioneering work of Raghavachari [20] on Hartree-Fock calculations for small Si clusters. For example, Raghavachari notes of one particular geometry of Si_7 that:

> "Another structure that we have considered is the edge-capped octahedron (7d). Though it is not expected to be a particularly stable structure, it was considered mainly to estimate the energy required to move the capping atom in 7c from one face to another. 7d can be considered as a *transition state* for such a process." [20] [emphasis added]

The details of the geometries 7c, 7d here are not particularly important, rather it is the concept that such geometries represent transition states that is relevant.

The ab-initio values for the clusters I calculated using the GAUSSIAN-98 software package; all cluster calculations were performed using the MPW1PW91 hybrid functional and the cc-pVTZ basis set. These ab-initio calculations alone represent some of the most intensive parts of my own research. Indeed, it is not uncommon to see entire journal articles devoted to the discussion of ab-initio calculations of small elemental clusters. Also, our choice of the all-electron MPW1PW91 level and the large cc-pVTZ basis set represent some of the most aggressive calculations feasible for small Si clusters; we are not aware of any published results for such clusters at this aggressive level.

It is perhaps equally important what strategies were avoided in choosing the cluster fitting properties. One example is the technique of using one level of theory to obtain the ab-initio geometries, and a different (higher) level of theory to obtain the energies. This is an entirely reasonable approach for projects that involve ab-initio calculations only. However, we are concerned that, for empirical modeling, it is more important to perform all the calculations at the same level of theory. Although there do not appear to be any comparative studies on which of these technique leads to the best empirical model, it is well-known that different levels of theory can introduce systematic differences in their calculated values, i.e. differences such as an overall shifting of the energies in some direction. We do not feel that is productive to attempt to force an empirical model to reproduce the systematic differences between two different types of ab-initio calculations.

A second example of a strategy that we deliberately avoid is fitting to forces. Even though such fitting would be expected to improve items such as transition states, we have become increasingly concerned about the effects of the small but nonzero differences between the calculated and reference values, which are always present in empirical modeling. For example, suppose that an ab-initio calculation has an equilibrium bond length of 2.20Å, and some force calculated slightly away from equilibrium at 2.30Å. However, suppose that our model (for a particular set of

parameters) has for the same cluster an equilibrium bond length of 2.25Å. Just what bond length are we supposed to calculate the force at? There are at least 3 reasonable options: we can use the "fixed" ab-initio value of 2.30Å, the "shifted" value of 2.35Å (i.e. 0.05Å past equilibrium), or the "percentage" value of around 2.352Å (i.e. around 4.5% past equilibrium). The differences in the calculated forces resulting from such arbitrary choices can be surprisingly large.

For the bulk properties we fit to both the energy curves of several crystalline phases as well as the band structure. The ab-initio calculations of bulk properties are not as problematic as those for clusters; for Si these reference values were taken from the older but well-established work of Cohen [19]. The results of for the band structure are shown in Figure 7. It is clear that the valence band is very well reproduced. The conduction band is seen to be more problematic, although this is also a known limitation of density functional theory, and of almost all existing empirical models. An overview of the reasons why ab-initio methods such as DFT are poor for the conduction band is given by Louie [31]. The problem is traced to the inability of the exchange-correlation energy to appropriately describe properties other than the ground state.

In Figure 8 we show the results of our model for several crystalline phases of Si, together with a comparison of our model to several other similar models. If we first consider the results in Figure 8 for only our model, we can see that the excellent agreement with the density-functional calculations. Of particular interest is the accuracy of the bcc and fcc phases which, as discussed in Section F, are generally though to be difficult to fit due to the large coordination numbers. These observations suggest the validity of the environment-dependent effects in our model. Perhaps even more remarkable is the comparison to other tight-binding models shown in Figure 8. The first three models are not environment-dependent, and it is perhaps not surprising that our results are an improvement. However, the model of Wang and Ho features an environment-dependent repulsive energy (as discussed in

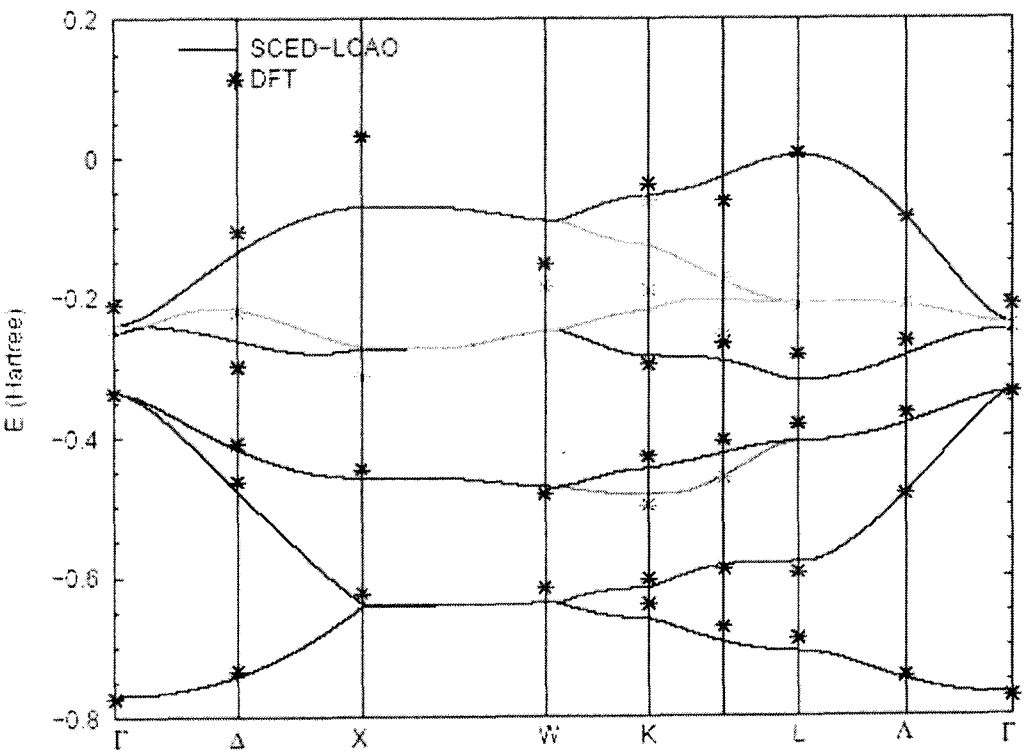

Figure 7. Results of our environment-dependent model for the band structure of (diamond) Si. The DFT values are taken from the work of Cohen in Ref. [19].

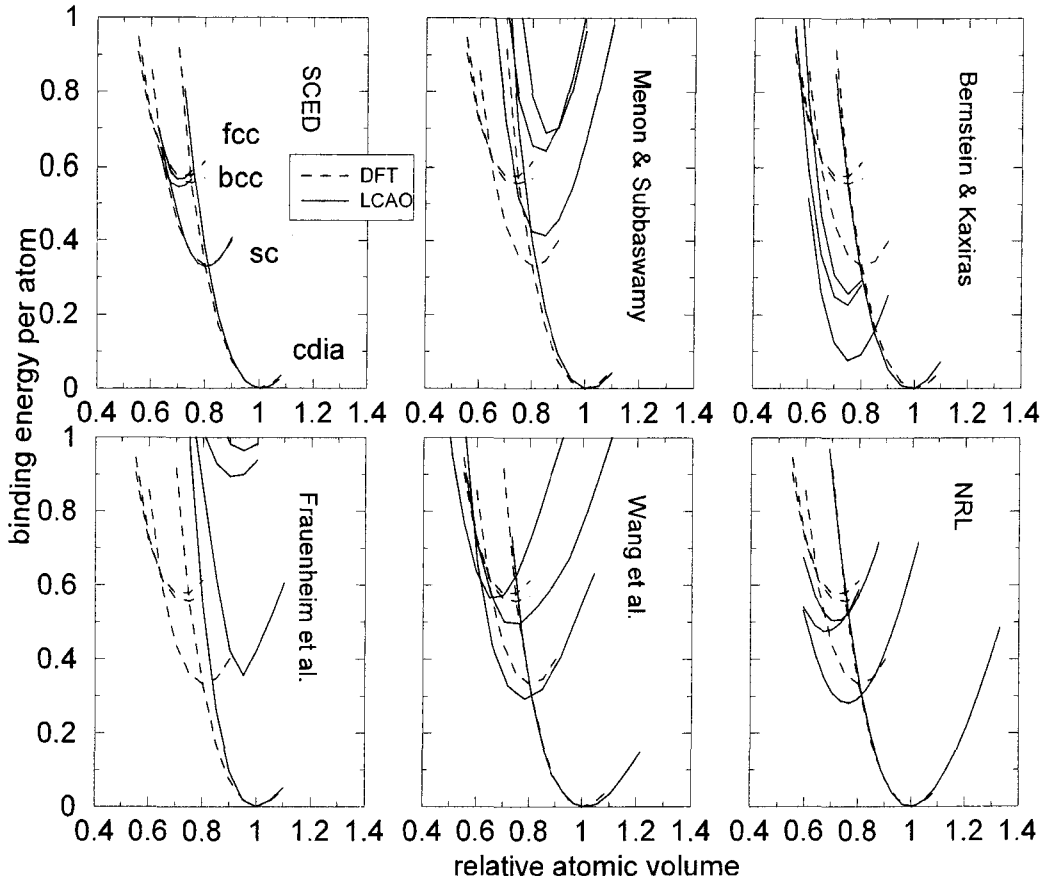

Figure 8. Results of our environment-dependent model for the bulk energy curves of Si, compared to the results of other empirical models. The DFT values are taken from the work of Cohen in Ref. [19] "Menon" refers to the model of Ref. [15]. "Kaxiras" refers to the model of Ref. [27]. "Frauenheim" refers the model of Ref. [10]. "Wang" refers to the model of Ref. [7]. "NRL" refers to the model of Ref. [28].

Section B), and the model of NRL features environment-dependent effects in the Hamiltonian, but without a treatment of self-consistency. The improvement over these models further validates the environment-dependent effects in our model, and offers evidence of the importance of treating the full iterative self-consistency problem.

H Applications for Si

In addition to developing an official set of parameters for Si, our research group has also applied this model to several interesting problems. My own part in the research group was more heavily oriented toward the development of the model. Most of this application work was done by a colleague, Dr. Ming Yu. As such this discussion will be brief, as it is intended here more to demonstrate the validity of my work on the model development of the model. One of the items that must be kept in mind is that no matter how carefully the least-squares fitting is done, a small residual does not necessarily indicate that model will be useful for applications. Conceptually, empirical modeling is a type of *extrapolation*, in that the parameters are adjusted by fitting them to the calculated properties of small systems, while the model is then used to study large systems. The "adjustment" of the parameters means that the fitting properties must be evaluated (very roughly) some 10^6 times, i.e. for 10^6 different sets of parameters. As such there is no way to put large systems in the fitting. One hopes that the extrapolation works, but in practice this must be tested, and it is these types of applications that validate the extrapolation.

Our first such application concerns the structure of the Si_{71} cluster. While bulk Si prefers a tetrahedral arrangement of atoms, most of the atoms of the Si_{71} cluster are on the exterior, and the "reconstruction" of such exterior atoms leads to complicated structures of low symmetry. Charge redistribution is of critical importance in such reconstructions, and it is with such systems that one might expect poor results from a model that does not properly account for charge transfer. In Figure 9 we show the structure of the cluster along with its pair distribution function. The pair distribution function gives the probability of finding an atom at a given distance from another atom. The results are compared to a density-functional-theory calculation of the same structure. Our calculation correctly reproduces the first and second nearest-neighbor peaks, demonstrating the ability of our method to reproduce the correct structural information as density

functional theory.

Our second and perhaps most important application is the reconstruction of the Si (001) surface. Starting with the ideal $P1 \times 1$ reconstruction, a molecular dynamics simulation was performed which resulted in the $C4 \times 2$ reconstruction, which is the experimentally observed reconstruction. This is shown in Figure 10. Two items are of particular interest for our discussion. First, this result occurs when the full self-consistent treatment of charge transfer is turned on, *but not when it is turned off.* When combined with our previous results, a pattern begins to emerge indicating that self-consistency is required in order to reproduce such results. Second, the combination of both speed and accuracy of our model allow it to break new ground in such calculations. Although the $C4 \times 2$ reconstruction can be obtained both by first-principles and other semi-empirical calculations, to the best of our knowledge ours is the only application in which it has been obtained entirely from the ideal $P1 \times 1$ reconstruction. Apparently, first-principles calculations are too slow, and other semi-empirical calculations are not accurate enough to obtain this result.

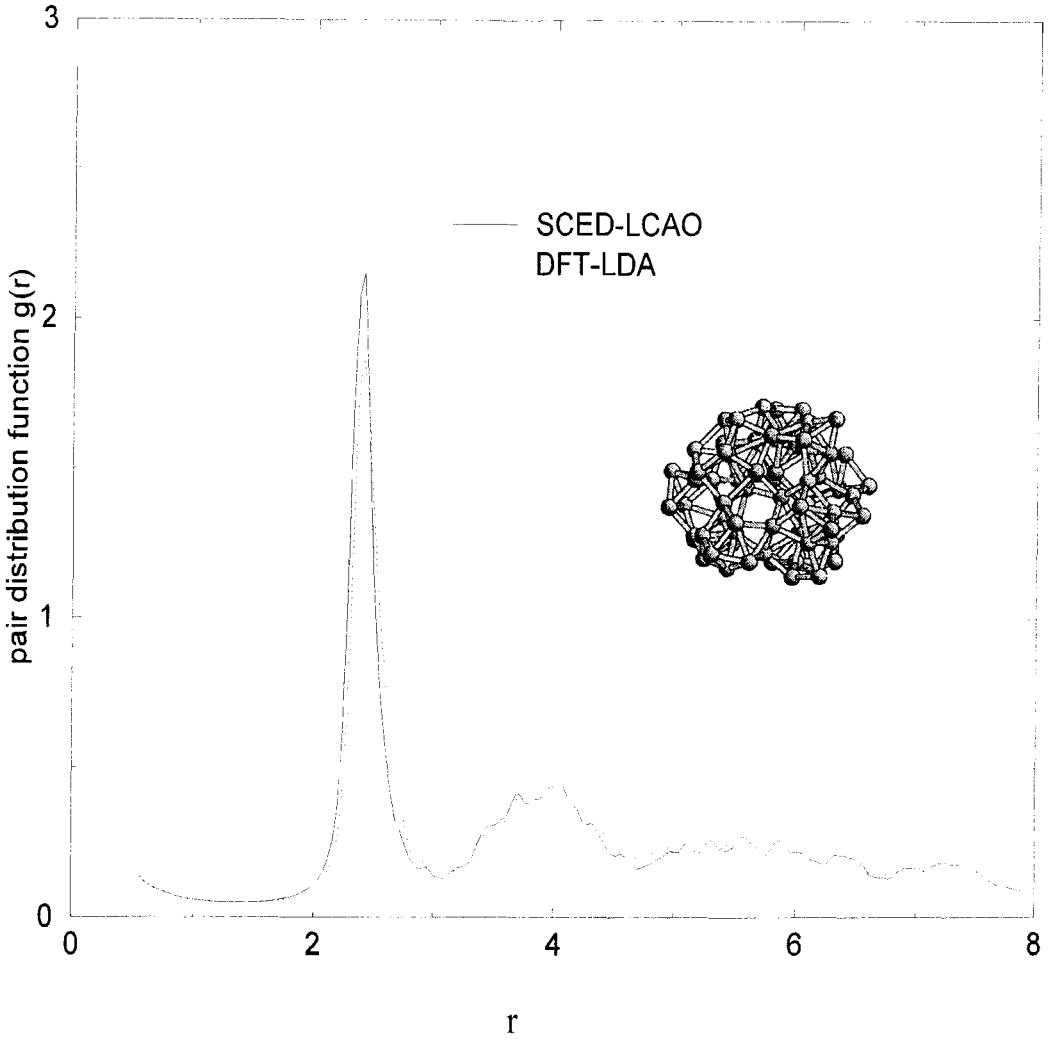

Figure 9. A comparison of the pair distribution function of the Si_{71} cluster calculated using our SCED-LCAO Hamiltonian and using a density-functional calculation from Ref. [32]. Also shown is the complicated low-symmetry structure of the cluster. The x-axis is in units of Angstrom.

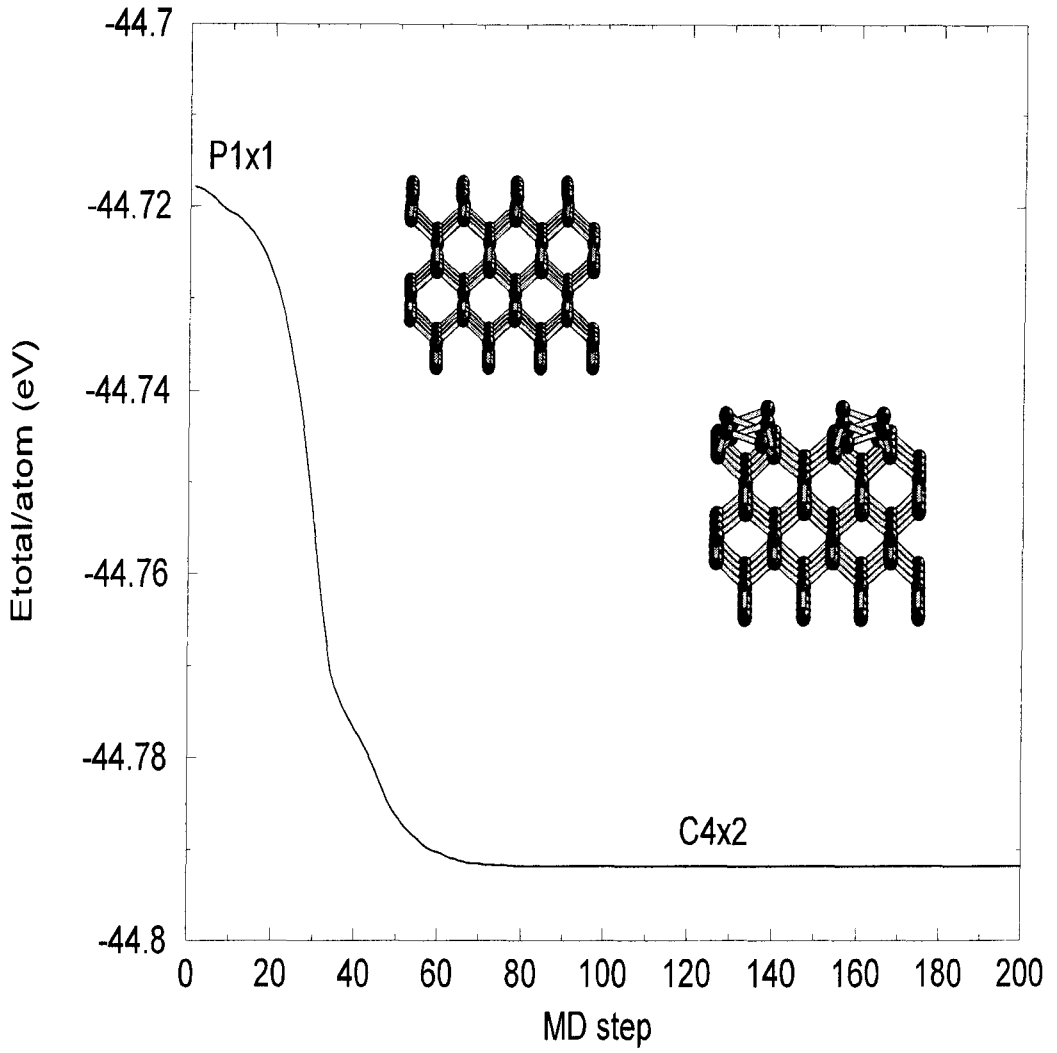

Figure 10. Total energy of the Si (100) surface as a function of the MD step, with one step corresponding to 2.5fs. Also shown are the original unrelaxed structure (left) and the relaxed C4x2 structure (right).

CHAPTER VII

CONCLUSION

In addition to the parameterization of our model for Si, we have also parameterized and applied our model to other group IV materials, most notably to carbon. In Figures 11 through 15 we show, mostly for reference purposes, the semi-empirical parameters for C, Si, and Ge, along with the calculated cluster and bulk properties for carbon and germanium. The parameters shown in Figure 11 follow the notation used in Ref. [30], which is slightly different from the notation used in out discussion; one should consult Ref. [30] if one is interested in using the parameters in Table 11. The same comments that apply to our results for Si also apply to these results for C and Ge. If there is an additional comment to be made about these results, it is that the chemistry of carbon is quite different from that of Si and Ge. So, while the success for Ge is perhaps less remarkable, the success for C further demonstrates the flexibility and transferability of our model.

In contrast to Si and Ge, carbon exhibits sp, sp^2, and sp^3 hybridizations. We have examined carbon clusters of various sizes, starting from various initial configurations, in order to examine the competition between these types of bonding, in determining the equilibrium structures of these molecules. In Figure 16 we show the "bucky-diamond" structure of C_{147}, which was obtained using our SCED-LCAO method. The interior of this structure has sp^3 bonding, while the exterior has sp^2 bonding. This type of structure has been previously obtained by first-principles calculations [35], but has not been obtained by other tight-binding calculations. This example demonstrates that our methodology is capable of capturing various bonding characteristics exhibited by carbon.

parameters	C	Si	Ge	parameters	C	Si	Ge
ε_s (eV)	-17.360	-13.550	-14.000	α_{sss} (Å^{-1})	2.153	3.036	2.956
ε_p (eV)	-8.329	-6.520	-6.850	d_{sss} (Å)	0.629	1.322	1.242
ε_s^* (eV)	-35.712	-13.428	-13.345	B_{sss} (Å^{-1})	-0.777	-0.746	-2.149
ε_p^* (eV)	-22.153	-7.908	-7.505	α_{sps} (Å^{-1})	2.013	2.176	2.317
α_s (Å^{-1})	-0.0329	0.247	0.257	d_{sps} (Å)	0.782	1.349	1.008
U (eV)	14.896	8.046	8.552	B_{sps} (Å^{-1})	-1.895	-0.796	-1.580
B_s (Å^{-1})	1.475	1.543	1.351	α_{pss} (Å^{-1})	1.881	2.349	1.878
A_p (eV)	-2.539	-1.257	-1.406	d_{pss} (Å)	0.377	2.026	1.248
B_p (Å^{-1})	-1.798	0.158	0.254	B_{pss} (Å^{-1})	0.236	-0.307	-0.157
α_p (Å^{-1})	3.115	2.739	2.176	α_{pps} (Å^{-1})	2.255	3.736	3.051
d_p (Å)	0.800	1.906	1.919	d_{pps} (Å)	0.547	2.282	2.015
B_{sp} (Å^{-1})	0.228	0.879	1.352	R_{cut} (Å)	4.0	6.5	6.5

Figure 11. Values of the semi-empirical parameters for C, Si, and Ge. The notation follows that of Ref. [30], which is slightly different from the notation used elsewhere in this report.

While we have also successfully obtained a parameterization for the heterogeneous system SiC, using an averaging technique as discussed previously, we have also encountered the limitations of the existing framework for heterogeneous systems consisting of elements from different groups, such as Li/Si. Therefore future research efforts are expected to focus on the radial function prototype, which is designed specifically to address systems consisting of several types of elements. In conclusion then, we have presented both a working model, and also a number of significant insights into the models themselves, for the simulation of large-scale systems using semi-empirical techniques.

Cluster	Symmetry	Present work	*ab initio* values[a]
C_2	$D_{\infty h}$	1.293 -5.228	1.244 -4.527
C_3	D_{3h}	1.329 -6.588	1.287 -6.586
C_3	C_{2v}	1.326 1.515 -5.988	1.256 1.459 -6.225
C_4	D_{4h}	1.488 -6.698	1.439 -6.746
C_4	D_{2h}	1.324 1.361 -6.520	1.288 1.306 -6.620
C_4	D_{2d}	1.382 1.554 -5.631	1.316 1.555 -5.566
C_4	T_d	1.577 -5.510	1.621 -4.830
C_5	D_{3h}	1.325 1.341 -7.124	1.277 1.282 -7.319
C_5	D_{5h}	1.487 2.113 -6.917	1.488 2.013 -6.578
C_5	C_{4v}	1.495 1.607 -6.547	1.443 1.668 -6.242
C_5	T_d	1.409 2.301 -5.521	1.417 2.314 -5.100

Figure 12. Results of our environment-dependent model for small C clusters. The ab-initio values were calculated using as in Table 1.

Cluster	Symmetry	Present work		*ab initio* values	
Ge_2	$D_{\infty h}$	2.326	Å	2.410	Å
		-2.284	eV	-2.294	eV
Ge_3	C_{2v}	2.440	Å	2.524	Å
		2.1287	Å	2.296	Å
		-3.073	eV	-3.172	eV
Ge_3	D_{3h}	2.237	Å	2.260	Å
		-3.129	eV	-3.116	eV
Ge_4	D_{2h}	2.402	Å	2.445	Å
		-3.678	eV	-3.834	eV
Ge_4	T_d	2.513	Å	2.625	Å
		-3.370	eV	-3.343	eV
Ge_4	D_{3h}	2.239	Å	2.252	Å
		2.258	Å	2.275	Å
		-3.017	eV	-3.027	eV
Ge_5	D_{3h}	2.348	Å	2.441	Å
		3.186	Å	3.253	Å
		-4.090	eV	-4.016	eV
Ge_5	C_{4v}	2.342	Å	2.424	Å
		2.547	Å	2.669	Å
		-3.888	eV	-3.815	eV
Ge_5	D_{3h}	2.218	Å	2.233	Å
		2.242	Å	2.234	Å
		-3.194	eV	-3.152	eV
Ge_5	T_d	2.297	Å	2.333	Å
		3.751	Å	3.809	Å
		-2.977	eV	-2.934	eV
Ge_6	D_{4h}	2.430	Å	2.518	Å
		2.786	Å	2.899	Å
		-4.158	eV	-4.191	eV
Ge_6	D_{3d}	2.395	Å	2.442	Å
		2.590	Å	2.484	Å
		-3.545	eV	-3.538	eV

Figure 13. Results of our environment-dependent model for small C clusters. The ab-initio values were calculated using as in Table 1.

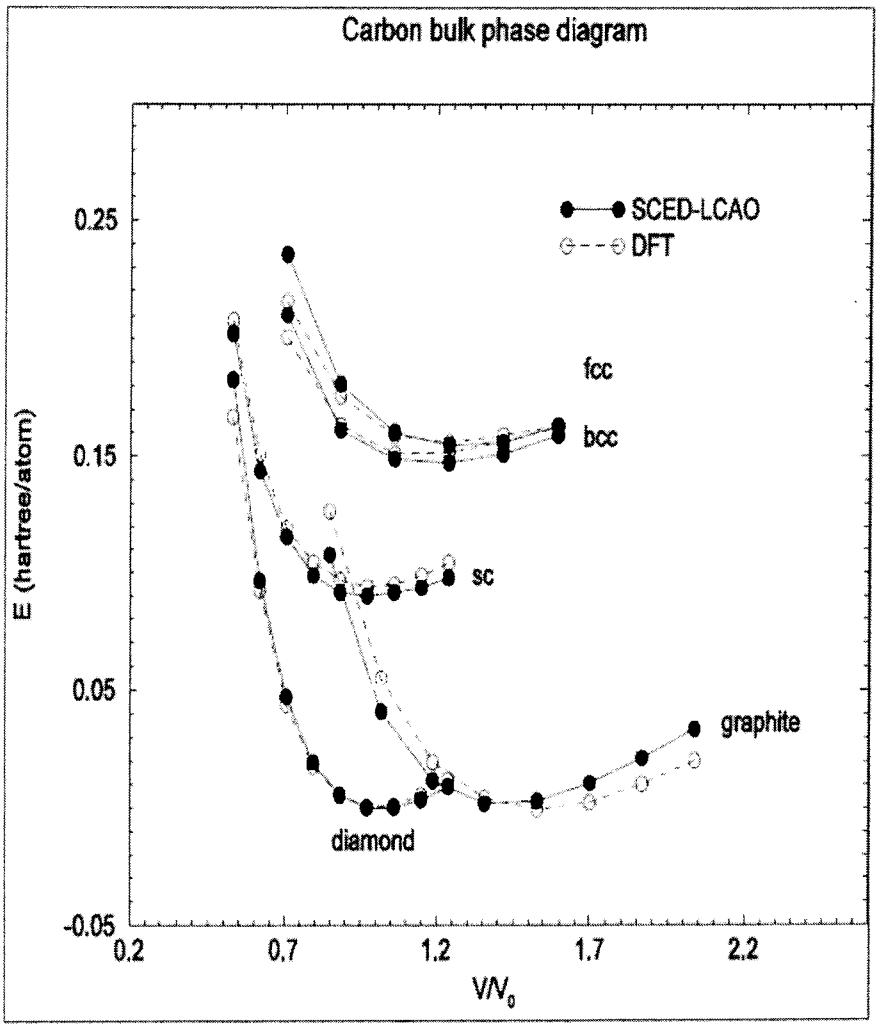

Figure 14. Results of our environment-dependent model for the bulk energy curves of C. DFT values are taken from Ref. [33], with the energy curve for graphite adjusted to match the experimental value from [34].

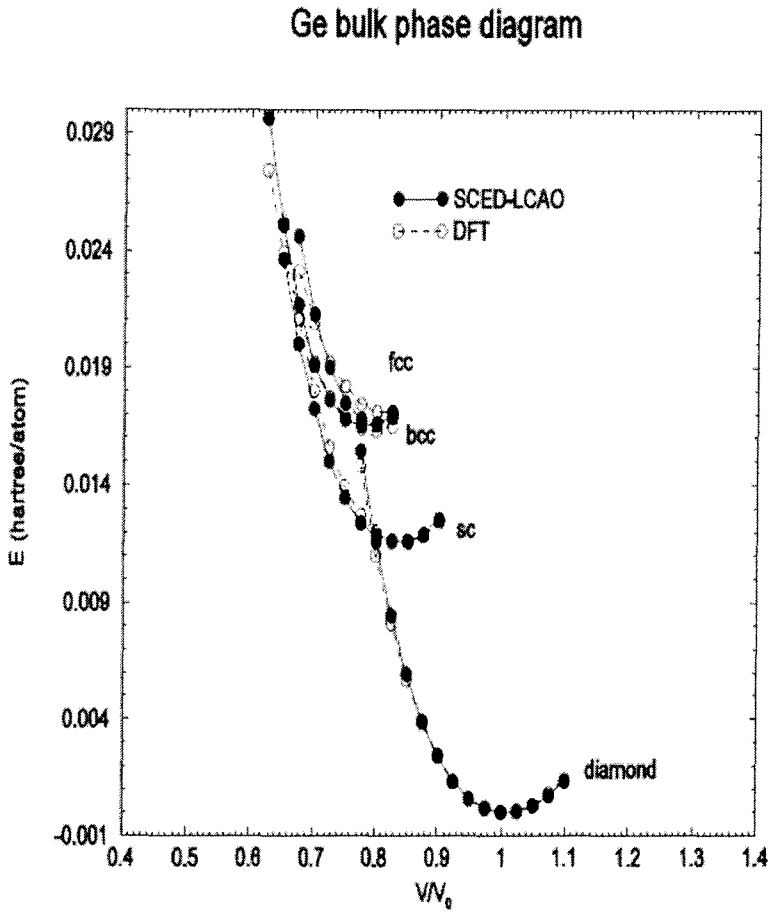

Figure 15. Results of our environment-dependent model for the bulk energy curves of Ge. DFT values are taken from Ref. [19]

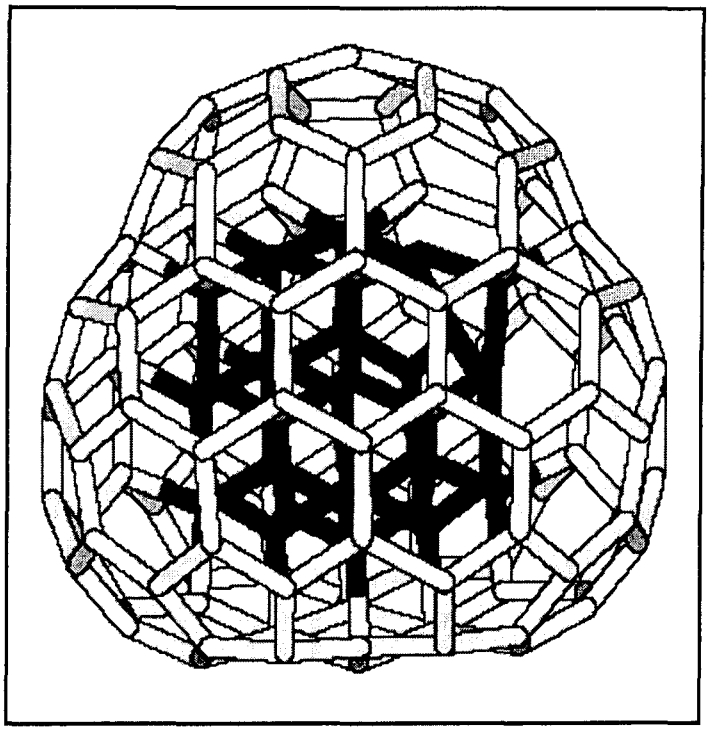

Figure 16. Equilibrium structure of the C_{147} "bucky-diamond" cluster, calculated using our semi-empirical model for carbon.

REFERENCES

[1] J.C. Slater and G.F. Koster, Physical Review **94**, 1498-1524, (1954).

[2] S. Froyen and W.A. Harrison, Physical Review B **20**, 2420-2422, (1979). See also W.A. Harrison, *Electronic Structure and the Properties of Solids* (New York, Dover, 1989).

[3] (There are a relatively large number of papers from the late 1970s; they are referenced in this later article) D.J. Chadi, Physical Review B **29**, 785-792, (1984).

[4] D.J. Chadi, Physical Review B **19**, 2074-2082, (1979).

[5] D. Tománek and M.A. Schülte, Physical Review Letters **56**, 1055-1058, (1986).

[6] P.B. Allen, J.Q. Broughton, and A.K. McMahan, Physical Review B **34**, 859-862, (1986).

[7] C.H. Xu, C.Z. Wang, and K.M. Ho, Journal of Physics: Condensed Matter **4**, 6047-6054 (1992).

[8] J.L. Mercer and M.Y. Chou, Physical Review B **47**, 9366-9376, (1993).

[9] D. Porezag, Th. Frauenheim, Th. Köhler, G. Seifert, and R. Kaschner, Physical Review B **51**, 12947-12957, (1995).

[10] Th. Frauenheim, F. Weich, Th. Köhler, S. Uhlmann, D. Porezag, and G. Seifert, Physical Review B **52**, 11492-11502, (1995).

[11] P.K. Sitch, Th. Frauenheim, and R. Jones, Journal of Physics Condensed Matter **8**, 6873-6888, (1996).

[12] S.J. Duclos, Y.K. Vohra, and A.L. Ruoff, Physical Review B **41**, 12021-12028, (1990).

[13] A.B. Anderson, Journal of Chemical Physics **62**, 1187-1188, (1975).

[14] R. Hoffmann, Journal of Chemical Physics **39**, 1397-1412, (1963).

[15] M. Menon and K.R. Subbaswamy, Physical Review B **55**, 9231-9234, (1997).

[16] W.M.C. Foulkes and R. Haydock, Physical Review B **39**, 12520-12536, (1989).

[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C* (Cambridge, Cambridge University Press, 1992).

[18] L.M. Canel, A.E. Carlsson, and P.A. Fedders, Physical Review B **48**, 10739-10750, (1993).

[19] M.T. Yin and M.L. Cohen, Physical Review Letters **45**, 1004-1007, (1980); M.T. Yin and M.L. Cohen, Physical Review B **26**, 5668-5687, (1982).

[20] K. Raghavachari, Journal of Chemical Physics **84**, 5672-5686, (1986); K. Raghavachari and J.S. Binkley, Journal of Chemical Physics **87**, 2191-2197, (1987); K. Raghavachari and C.M. Rohlfing, Journal of Chemical Physics **89**, 2219-2234, (1988).

[21] J. Cioslowski, Journal of the American Chemical Society **111**, 8333-8336, (1989).

[22] J.L. Mercer and M.Y. Chou, Physical Review B **49**, 8506-8509, (1994).

[23] L. Velázquez, G.N. Phillips, R.A. Tapia, and Y. Zhang, Computational Optimization and Applications **20**, 299-315, (2001).

[24] A. Laio and M. Parrinello, Proceedings of the National Academy of Sciences (PNAS) **99**, 12562-12566, (2002).

[25] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Philadelphia, SIAM, 1996).

[26] H.H. Rosenbrock, The Computer Journal **3**, 175-184, (1960).

[27] N. Bernstein and E. Kaxiras, Physical Review B **56**, 10488-10496, (1997).

[28] M.J. Mehl and D.A. Papaconstantopoulos, Physical Review B **54**, 4519-4530, (1996).

[29] See for example F.J. García-Vidal, A. Martín-Rodero, F. Flores, J. Ortega, and R. Pérez, Physical Review B **44**, 11412-11431, (1991).

[30] C. Leahy, M. Yu, C.S. Jayanthi, and S.Y. Wu, Physical Review B **74**, 155408, 1-13 (2006).

[31] M.S. Hybertsen and S.G. Louie, Physical Review B **34**, 5390-5413 (1986). See also J.P. Perdew and M. Levy, Physical Review Letters **51**, 1884-1887 (1983); L.J. Sham and M. Schlüter, Physical Review Letters **51**, 1888-1891 (1983).

[32] O.F. Sankey and D.J. Niklewski, Physical Review B **40**, 3979, (1989); A.A. Demkov, J. Ortega, O.F. Sankey, and M.P. Grumbach, Physical Review B **52**, 1618, (1995).

[33] C. Mailhiot and A. K. McMahan, Physical Review B **44**, 11578-11591 (1991).

[34] M.T. Yin and M.L. Cohen, Physical Review B **29**, 6996-6998 (1984).

[35] J.Y. Raty, G. Galli, C. Bostedt, T.W. van Buuren, and L.J. Terminello, Physical Review Letters **90**, 037401, 1-4 (2003).

CURRICULUM VITAE

NAME: Chris Leahy

ADDRESS: Department of Physics
University of Louisville
Louisville, KY 40292

EDUCATION: M.S. Physics
University of Louisville
1998