Chapter 4

Inversion of Cohesive Energy Curves

> The physicist cannot ask of the analyst to reveal to him a new truth; the latter could at most only aid him to foresee it.
>
> – Henri Poincaré [89]

Having gained insight into interatomic forces mediated by hybrid covalent bonds in ideal lattices, we now ask what truths may be revealed concerning global trends across more complex structures with different bonding character. Quantum approximations are very useful in suggesting qualitative trends and providing physical understanding, but one wonders whether any quantitative information can be extracted directly from *ab initio* energy data without resorting to the uncontrolled and inconclusive fitting approach. Inspired by Poincaré, we may see if pure mathematics can lead us in a fruitful direction.

So, what is the basic mathematical problem we are interested in solving? The answer is, of course, that we wish to reproduce the many-dimensional Born-Oppenheimer energy surface with a relatively simple functional form. Thus stated, the inverse problem is incredibly overdetermined, and we must settle for an approximate solution obtained by some sort of optimization procedure. However, if the dimensionality of the manifold we wish to fit is sufficiently reduced, then we may be left with a nonsingular and tractable inverse problem. For example, we may hope to uniquely determine a force law containing a single, one-variable, continuous function from a one-parameter energy curve.

In 1980, Carlson, Gelatt and Ehrenreich (CGE) showed that this is indeed possible by proving an inversion formula which gives the pair potential that exactly reproduces a given cohesive energy versus volume curve [90]. In spite of its mathematical elegance, the CGE formula has so far not produced potentials of practical use or been connected with theories of chemical bonding, and hence it has only been employed by a handful of authors. Nevertheless, it is such a radically different and aesthetically appealing approach compared to brute-force fitting, that in this chapter we set out to understand its limitations and extend its applicability to more realistic functional forms for covalent solids. In order to make progress toward these goals, it will be necessary to invent a new way to think about the mathematics of inversion. Following this work, fitting will still play the central role in developing potentials because the important regions of the Born-Oppenheimer surface are too vast to permit an exact solution, but inversion will at least provide sorely needed guidance for the fitting process and build our physical intuition.

4.1 Pair Potentials

4.1.1 The Carlsson-Gelatt-Ehrenreich Formula

We begin with the original derivation of the CGE formula [90]. Consider an isotropic crystal structure, for which the set of displacement vectors from one atom to all others is the same for every atom in the crystal, up to trivial rotation and inversion symmetry operations[1]. Let $\{\vec{R}_i\}$ denote the set of atomic positions about a central atom located at the origin. Let r be the nearest neighbor distance, and group the atoms into shells of radius $s_p r$ containing n_p atoms each. Number the shells so that $s_1 < s_2 < s_3 < \ldots$. By definition $s_1 = 1$. Uniform expansion (dilation) of the crystal is described by varying r while keeping the structural quantities $\{s_p\}$ and $\{n_p\}$ constant. For simplicity, assume that the cohesive energy is completely described by a pairwise, radial interaction,

$$E[\phi](r) = \sum_i \phi(R_i) = \sum_{p=1}^{\infty} n_p \phi(s_p r), \tag{4.1}$$

a condition imposed by CGE that we shall eventually relax. Define a weighted scale transformation operator T_p whose action on a function ψ is given by,

$$T_p \psi(x) = n_p \psi(s_p x). \tag{4.2}$$

Note that T_1 amounts to multiplication by the constant n_1 (since $s_1 = 1$), so $T_1^{-1} = 1/n_1$. The derivation proceeds by expressing the energy $E(r)$ (a function of first neighbor distance) as the result of a *linear* operator acting on the pair potential $\phi(r)$ (a function of atomic separations or shell radii),

$$\begin{aligned} E(r) &= \sum_{p=1}^{\infty} T_p \phi(r) \\ &= \left[T_1 \left(1 + \sum_{p=2}^{\infty} T_1^{-1} T_p \right) \right] \phi(r). \end{aligned} \tag{4.3}$$

If we view the operator in parentheses as $(1 + U)$, then its inverse (if it exists) would be given by the Neumann (geometric) series formula [91], $1 - U + U^2 - U^3 + \ldots$. Letting Θ denote the full operator in brackets, we thus have a formal expression for its inverse,

$$\Theta^{-1} = \left(1 - \sum_{p=2}^{\infty} T_1^{-1} T_p + \sum_{p=2}^{\infty} \sum_{q=2}^{\infty} T_1^{-1} T_p T_1^{-1} T_q - \ldots \right) T_1^{-1}, \tag{4.4}$$

which leads to the desired inversion formula,

$$\phi[E](r) = \frac{1}{n_1} E(r) - \sum_{p=2}^{\infty} \frac{n_p}{n_1^2} E(s_p r) + \sum_{p,q=2}^{\infty} \frac{n_p n_q}{n_1^3} E(s_p s_q r) - \ldots. \tag{4.5}$$

[1] This ubiquitous assumption is not required, as described at end of the next section.

For mathematical rigor, we should properly state the conditions on $E(r)$ that guarantee the existence of an inverse and the convergence of the series in Eq. (4.5). A sufficient condition for convergence is $|E(r)| < A/r^3$ for some $A > 0$ in three dimensions [90]. As a special case, if the energy has a finite cutoff a, *i. e.* $E(r) = 0$ for $r > a$, then the series trivially converges (and is finite).

As long as $E(r)$ is a well-behaved function (like most encountered in physics), problems with convergence can only come from a slowly decaying tail. For a long-range force, like the Coulomb force in ionic crystals, it is well known that our inverse problem is ill-posed because the series in Eq. (4.1) is only conditionally convergent, meaning that its value depends sensitively on the order of summation [92]. For a covalent (or metallic) solid, however, we should not worry too much about convergence because, as we shall see shortly, the long-range tails of *ab initio* energy curves and inverted potentials should not be taken very seriously.

4.1.2 The Chen-Möbius Theorem

In 1990, Chen proved an inversion theorem [93] that generalizes the Möbius inversion formula of number theory [94, 95] from discrete (integer) to continuous (real) variables. (See Appendix C for a statement of the Möbius theorem.) Chen's result is that if,

$$F(x) = \sum_{n=1}^{\infty} f(nx), \tag{4.6}$$

then

$$f(x) = \sum_{n=1}^{\infty} \mu(n) F(nx), \tag{4.7}$$

provided the sums converge, where the inversion coefficient is simply the Möbius function, defined by,

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ (-1)^{\ell} & \text{if n = product of } \ell \text{ distinct primes} \\ 0 & \text{otherwise} \end{cases} . \tag{4.8}$$

The aesthetic appeal of the Chen-Möbius (CM) theorem is that it shows how in certain special cases (of high lattice symmetry) the complicated multiple sums in the CGE formula combine into a single sum, indicating subtle cancelations between the oscillating terms. Chen's result has generated some excitement as a practical application of number theory, a field of pure mathematics that rarely finds applications in physics [96]. Although its physical accuracy remains to be seen, it provides an elegant formal approach to solve a wide range of important inverse problems in physics (*e. g.* getting the phonon density of states from the specific heat of a solid versus temperature) [93]. A natural application of the CM formula is to the pair potential inversion problem of Eq. (4.1) just described. As stated, the theorem already solves the problem for the case of a one-dimensional, equally-spaced lattice (since $s_p = p$ in that case). With this approach, Chen and others have rederived the CGE formula for various, special crystal lattices (square and hexagonal planes [97], diamond [98]), and this year they have presented a general Möbius formulation of the CGE formula for multi-dimensional lattices [99]. For a variety of crystal structures, it is possible to derive pair potential inversion formulae of the form,

$$\phi(r) = \sum_{n=1}^{\infty} \tilde{\mu}(n) E(\tilde{s}_n r), \tag{4.9}$$

where $\tilde{\mu}(n)$ is a generalized Möbius function and $\{\tilde{s}_n\}$ is the closure of $\{s_n\}$ under multiplication [99]. A drawback of the CM approach, however, is that the quantities $\tilde{\mu}(p)$ must be recalculated for every crystal structure, which is a nontrivial task except in highly symmetric lattices.

It is also restrictive to base the method on crystal lattices because it is possible to apply the CGE formula to *any* periodic structure, even a large super-cell of amorphous material. Although no author has reported such a result, the insight is that a solid with only pairwise, radial interactions is simply a collection of bonds (atomic separations), which can be divided into "shells" of increasing size as above, with the caveat that bonds in a shell may not correspond to the sequence of neighbor radii about a typical atom. The topology and geometry of atomic arrangements are irrelevant for total energy versus volume curves. This generalization to complex crystal structures and defects is difficult to make with the CM approach, since the generalized Möbius coefficients are very complicated for low symmetry structures.

4.1.3 Limitations of the CGE Formula

Although the CGE formula is a significant innovation in materials theory, it is only the first step toward an inversion method of practical use, owing to several serious limitations. First, it can only be applied to cohesive energy versus volume curves, thus excluding important chemical bonding changes that occur under shear strains and internal rearrangements. This is a major drawback for covalent solids because the subtleties of covalent bonding mostly arise from nonuniform lattice distortions. It is straightforward to theoretically estimate the overall cohesive energy as a function of volume [76], but the small changes in total energy responsible for interatomic forces in a bulk solid at the equilibrium volume are much harder to calculate. It would be desirable to have an inversion method capable of extracting interatomic forces from *ab initio* shear strain data.

A second limitation is that of functional form. Clearly, a pair potential alone cannot hope to give a reasonable description of bonding in covalent solids (although it might adequately describe volume effects with all angles fixed). Mathematically, the CGE formula is hard to generalize, because the derivations of CGE and Chen both rely on the linearity of the energy functional $E[\phi]$. A many-body interaction always contains some nonlinear combination of the unknown functions to be obtained by inversion. Efforts have been made [100] to include many-body interactions using the CGE formula within the N-Body Potential format of Finnis and Sinclair, which mimics the second moment approximation of TB models [101]. This functional form is the simplest example of a many-body interaction, being a nonlinear combination of sums of pairwise, radial terms with no explicit angular dependence. Because the CGE formula does not handle nonlinearity, the authors of this study find it necessary to include various *ad hoc* assumptions and supplementary experimental inputs (like the elastic constants and the vacancy formation energy), which is contrary to the motivation for doing inversion in the first place: to extract parameter–free potentials directly from *ab initio* data. With the CGE and Chen-Möbius formulae, it appears to be impossible to perform even the simplest many-body inversion directly from *ab initio* cohesive energy curves, and there is no obvious extension to handle angular forces. A new approach is clearly needed to perform meaningful inversions for covalent solids.

The third, and perhaps most serious, limitation of the CGE and CM approaches is that the formal mathematics obscures the physical meaning of the inversion process. Although the inversion is mathematically exact, it is difficult to assess the physical validity of the resulting interatomic potentials. Unlike number theorists who are mostly concerned with mathematical rigor, physicists must always question the physical relevance of formally exact solutions, which often overextend the validity of simple models. In the following sections, we shall see how all of these limitations can be removed with a deceptively simple trick.

4.1.4 Recursive Inversion

We now present a recursive proof of the CGE formula that can be naturally extended to much more complicated situations[2]. The idea is very simple: separate the first shell term from the sum in Eq.(4.1), and solve for $\phi(r)$,

$$E(r) = n_1\phi(r) + \sum_{p=2}^{\infty} n_p\phi(s_p r) \tag{4.10}$$

[2]The same trick can also be applied to the Möbius theorem itself, as described in Appendix C, leading to an alternative version of the Möbius inversion formula and some interesting connections between combinatorics and number theory.

$$\phi(r) = \frac{1}{n_1}\left(E(r) - \sum_{p=2}^{\infty} n_p \phi(s_p r)\right), \tag{4.11}$$

The CGE formula follows by recursive substitution. The original derivation of the inversion formula by CGE relies on the linearity of the functional $E[\phi]$, and thus cannot be generalized to higher orders of cluster expansion, in which products or powers of radial functions appear. All that is required for this derivation, however, is the ability to solve for $\phi(r)$ in terms of $\phi(s_p r)$ for $p \geq 2$, which permits a straightforward generalization to higher order terms (and with a little more thought, even shear strains).

The recursive approach also reveals the mathematical structure of the CGE formula in a simple manner: the pair potential at r is chosen so that the first neighbor contribution to the cohesive energy, $n_1\phi(r)$, provides exactly the energy left over from interactions with higher shells. A simple consequence of this observation is that, if $\phi(r')$ is known for all $r' > r$, then Eq.(4.11) uniquely determines $\phi(r)$. This suggests an analytic procedure that does not involve an explicit formula like Eq.(4.5). Suppose that the potential has a cutoff distance a such that $\phi(r) = 0$ for $r > a$. The pair potential can then be generated by solving for $\phi(r)$ using Eq.(4.11) in order of decreasing r starting at the cutoff. All the complicated sums in the CGE formula are implicitly contained in the procedure. In addition to providing a simpler way to compute the potential, the recursive approach is crucial for nonlinear energy functionals in which it would be cumbersome even to write down explicit formulae.

4.1.5 *Ab Initio* Cohesive Energy Curves for Silicon Crystals

In order to explore the applicability of the inversion procedure to covalent solids, we shall consider cohesive energy curves for a set of seven silicon crystal structures, chosen to represent all the important local bonding states of bulk material. Of course, the set must include the diamond structure (Si-I) for sp^3 hybrid bonds. The BC8 structure (Si-III), experimentally observed upon relaxation from high pressure, which also has coordination four, is included to represent distorted sp^3 hybrids [103]. The six-fold

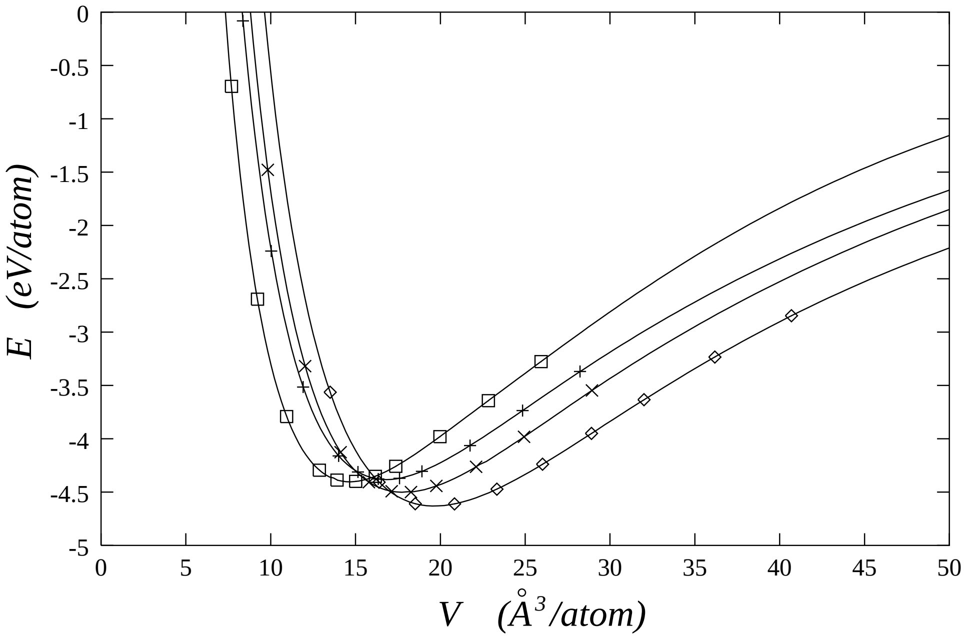


Figure 4.1: Interpolation of *ab initio* cohesive energy versus volume data for the low energy silicon crystal structures: cubic diamond (diamonds), BC8 ($\times$), BCT5 ($+$) and β-tin (squares).

coordinated β-tin (Si-II) phase, the first experimental high-pressure phase transition from diamond, is chosen to model low-energy metallic bonds [104]. The hypothetical BCT5 structure [105], a five-fold coordinated lattice predicted by *ab initio* calculations to be low in energy, is included because it contains an interesting mix of intra-planar metallic bonds and inter-planar covalent bonds, four and one per atom, respectively [106]. The hypothetical simple cubic (SC) and face-centered cubic (FCC) crystals, with coordinations six and twelve, respectively, are chosen to examine the canonical metallic arrangements. Finally, the hypothetical three-fold coordinated graphitic lattice is included for local sp^3 hybridization. In order to capture the planar nature of graphitic bonds, energies are computed with c fixed, but for all other crystals we compute energy versus volume with all angles fixed.

The total energy of each structure is calculated with density functional techniques [2] in the local density approximation (LDA) [3], using a plane-wave basis (12 Rydberg cutoff) and 512 k-points in the full Brillouin zone. By using *ab initio* computational methods, we can generate a wide range of reliable energy data for structures that are not experimentally accessible, including nonequilibrium volumes and exotic crystal lattices. Of course, these methods are validated by close agreement with experiment whenever comparison is possible [4]. Density functional methods have been particularly successful for covalent solids like silicon. The predicted lattice parameters and elastic constants are usually within a few percent of experimental values (although the former tend to be somewhat low with LDA). Predicted energy differences are also very accurate, but LDA binding energies are differ with experimental values by as much as 50% due to the well-known difficulty of LDA to represent accurately the energies of isolated atoms [102]. To avoid this problem, only atomic volumes smaller than $(3.54 \text{ Å})^3$ are used in constructing cohesive energy curves. At this volume, covalent bonds have been destroyed and almost half the binding energy has been lost. The cutoff in volume is chosen after a careful analysis of wider ranges of LDA data for all the crystals. Once silicon crystals are expanded to this volume, both the pressure dE/dV and its derivative have saturated, and the inflection point of each curve has been passed. Beyond this volume, the LDA data is considered unreliable and is replaced with an interpolant.

Following CGE, we use rational interpolation in the region of calculated cohesive energy values and an exponential tail, $a \exp(-br - cr^2)$, for larger distances. The coefficients a, b and c are chosen so that the interpolant is continuous with two continuous derivatives. The interpolated cohesive energy curves and LDA data points are shown in Fig. 4.1 for the important low energy structures, in close agreement with previously published results [104, 103, 102, 106]. Note that the covalent structures, diamond and BC8, have larger volumes, due to directed bonding, than the metallic β-tin phase, with the mixed-bonding BCT5 phase in between. The first high-pressure phase transition in silicon, predicted by a tangent construction, is from diamond to β-tin [104].

4.1.6 Physical Validity

The first question one should ask concerning an exact inversion procedure is whether the resulting potential is unique or instead depends somehow on the choice of input. It is customary in the inversion literature to refer to "the inverted (or *ab initio*) potential" for a particular material, but before this work no one has considered inversion of multiple cohesive energy curves for the same material. We shall see that the common terminology is misleading, due to significant sensitivity to the input data and the details of the inversion procedure. Our goal shall be to produce "an *ab initio* potential" with optimal physical validity.

The inversion procedure is implemented by starting at a large cutoff of 7.0 Å, where the cohesive energy is essentially zero (less than 0.001 eV/atom), and solving for $\phi(r)$ at equally spaced mesh points ($\delta r = 0.011$Å) using piecewise quadratic interpolation to evaluate $\phi(r')$ for $r' > r$. Fig. 4.2 shows the set of pair potentials that result from inverting the our silicon cohesive energy curves. The wide variation in these curves clearly demonstrates the nonuniqueness of inverted potentials.

Now that we have verified the multiplicity, however, we must ask another, more important question: is any one inverted potential physically meaningful, even for the structure from which it was derived? For example, we would hope that the diamond pair potential could at least provide a reasonable description of the diamond phase. To answer this question, consider the most striking feature of all the potentials (including the diamond potential): there is the strong repulsion ($d\phi/dr < 0$) at the first neighbor distance in the diamond structure. This means that the equilibrium spacing is set by a balance between first-neighbor repulsion and weak attractions from second, third and fourth neighbors, which contradicts our theoretical understanding of covalent bonds. It also inconsistent with the elastic constant analysis of Chapter 3. Similar results have

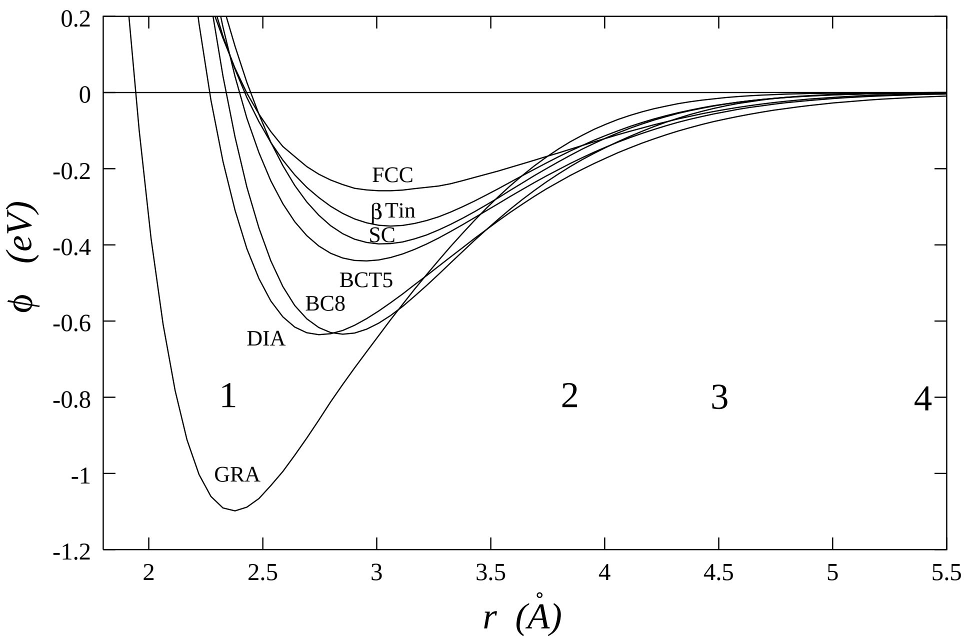


Figure 4.2: Pair potentials for silicon obtained from exact inversion of the raw *ab initio* cohesive energy versus volume curves for seven experimentally-observed (DIA, BC8, β-Tin) and hypothetical (GRA, BCT5, SC, FCC) crystal structures. Numbers indicate the positions of the first four neighbor shells in the ground-state diamond lattice.

been obtained by Wang *et. al.* for Si, C and SiC, but they were unable to explain the problems. Even in their pioneering work, CGE found unphysically long-range tails and strong first-neighbor repulsions for metals K, Cu and Mo. In fact, no physically reasonable potential has yet been produced for silicon (or any other material) using the CGE formula without modifications, and all inverted potentials share the problem of artificially long range.

These problems, which are difficult to see in the CGE and CM formulae, are transparent in the recursive inversion formula. Because the solution begins at a large cutoff and proceeds to smaller distances, the tail of the inverted potential comes from the cohesive energy of a greatly expanded crystal whose first neighbors are near the cutoff, which is exactly the interaction between isolated atoms in the gas phase[3]. This tail is then used to describe interactions with higher shells when determining the potential at the nearest neighbor distance in the equilibrium solid, so that inaccuracies in the long-range tail are magnified and propagated to smaller separations. The problem is that long-range interactions in a solid are screened compared to isolated atoms at the same separation, with the effect being greatest at large distances. Bare (negligible wavefunction overlap) atomic interactions are known to fall off as power laws (like r^{-6} for Van der Waals dipole-dipole correlation forces) [107]. On the other hand, screened cohesive forces in covalent solids are likely to have exponential decay, like the electron-screened ion-ion interaction [92], which is consistent with almost every empirical potential for silicon and related materials [19]. In summary, although the inversion procedure is mathematically exact, it does not produce realistic potentials because it requires the assumed functional form to be valid over the entire range of atomic volumes from solid to gas.

[3]This assertion is supported by Fig. 4 of Ref [17] in which the tail of the inverted pair potential for solid copper is seen to overlap perfectly with the binding energy curve of the Cu_2 molecule.

4.1.7 An Essential Modification

Before giving up on inversion, we can can try to modify the procedure to rectify the problems. The essential change is to forgo the requirement that an inverted potential exactly reproduce an *entire* cohesive energy curve. Instead, let us focus on condensed volumes typical of solid and liquid environments, whose exact energies can be preserved with any choice of tail for the potential. Ideally, we would take the tail of the potential from theoretical calculations, and proceed to smaller separations with the recursion procedure. Since reliable theory for long-range forces is not available, however, the next best thing is to experiment with various choices for the tail and see if any of them give reasonable results. Altering the tail of an interatomic potential implies a corresponding change in the tail of the cohesive energy curve, which can be modeled by multiplying the energy by a cutoff function that is unity for small first neighbor distances r with gentle decay to a cutoff distance $r = a$ over a range $\Delta r = \delta$. Modifying the long-range tails of cohesive energy curves does not destroy much meaningful *ab initio* data. Recall that due to the problems in treating isolated atoms with LDA, tails of energy curves always come from *ad hoc* extrapolation from the range of valid LDA data.

After exploring a number of possible cutoff functions, the following turned out to be most useful[4],

$$f_c(r) = \begin{cases} 1 & \text{if } x \leq 0 \\ \exp(\sigma)\exp\left(\frac{\sigma}{x^2-1}\right) & \text{if } 0 < x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}, \tag{4.12}$$

where $x = (r-(a-\delta))/\delta$. This choice of cutoff function has all derivatives continuous at $r = a$, which is important for numerical stability, as described in the next section. The parameters a and σ control the range of the potential. Experience in adjusting these parameters shows that it is important to keep interactions with second (and higher)

[4] The improved cutoff function in Eq. (5.2), which has two continuous derivatives, was discovered after the completion of the inversion study. If that function were used here, the slight shoulder in the inverted pair potentials of Fig. 4.3 would be smoothed out.

neighbors small for silicon, which is consistent with the theoretical arguments of the previous section. A reasonable choice is to set $a = a_{SW} = 3.77118$ Å, the SW cutoff distance, and $\sigma = 2\sigma_{SW}/\delta = 3.49183$ Å, which gives the inverted potentials exactly the SW asymptotic dependence at the cutoff. Although the cutoff is just short of the equilibrium second neighbor distance in the diamond structure, other crystals have multiple equilibrium neighbor shells inside the cutoff. The smoothing range $\delta = 1.2$ Å is chosen to allow for flexibility in cutting off the original curve while maintaining the exact energy values near the minimum in order to preserve important equilibrium properties (*e. g.* binding energy, lattice constant, and bulk modulus). *Ab initio* energies are not disturbed within 10% of the equilibrium bond length, where covalent bonds are well–defined.

The effect of imposing such a cutoff is illustrated in Fig. 4.3 for the case of the diamond lattice. Note that the modified potential has a deep minimum at the first neighbor distance, and closely resembles the fitted SW pair potential. At this point it may seem like we have made some arbitrary choices, thus tarnishing the aesthetic appeal of inversion, but they are validated *a posteriori* by the remarkable agreement between inverted potentials and well-known theories of covalent bonding. It is also important to keep mind that all energies at typical condensed volumes are still *exactly* reproduced; energies are only compromised at much larger (gaseous) volumes, where the data is not so reliable anyway.

4.1.8 Numerical Stability

Although the inversion procedure is exact in the sense that the input energy curve (after imposing the cutoff) is perfectly reproduced, numerical instability can cause artificial and unphysical oscillations in the inverted curve. The problem is that the inversion summation becomes numerically unstable when the set of scaled separations $\{s_p\}$ is very closely spaced (just above 1.0) for small p. In such cases, the CGE formula involves the

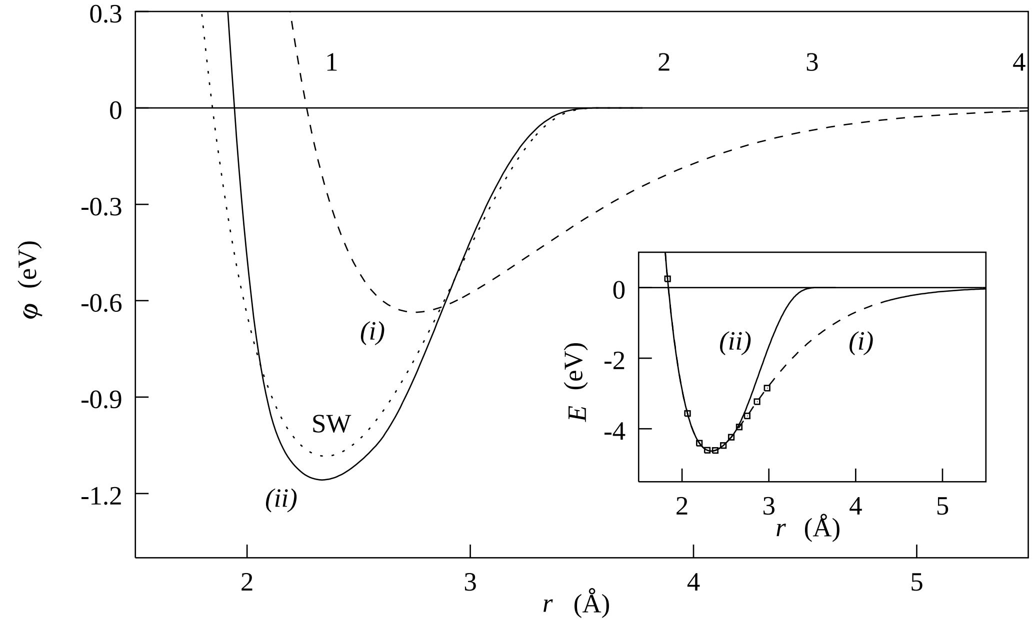


Figure 4.3: The inverted pair potential for silicon in the diamond phase (i) before and (ii) after the cutoff is imposed, compared with $\phi_{SW}(r)$ (dotted line). Numbers inside the figure indicate shell radii in the diamond lattice. The inset shows the diamond LDA data and the interpolant (i) before and (ii) after imposing a cutoff.

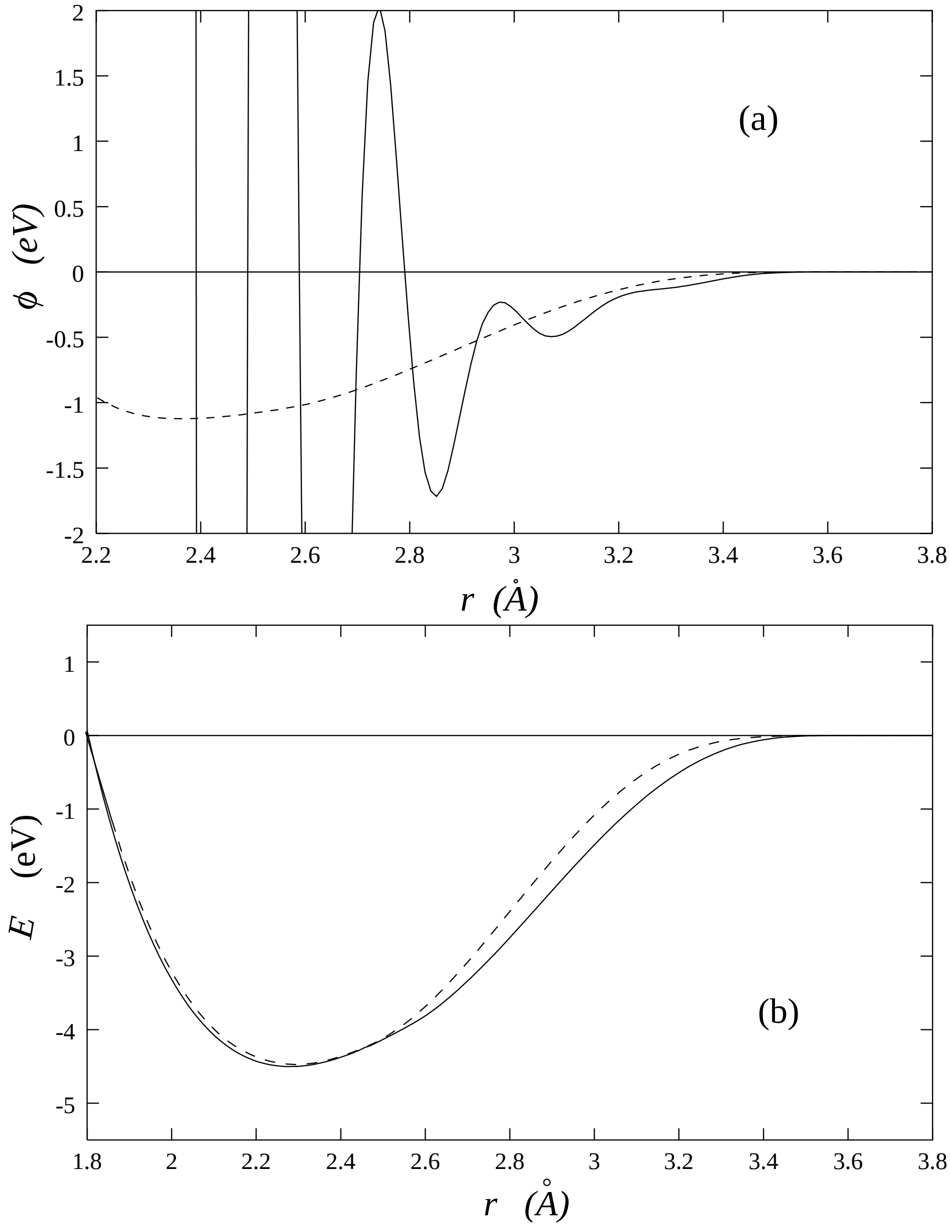


Figure 4.4: Removal of numerical instability of the CGE formula for the BC8 phase of silicon. In (a) are shown the inverted pair potential for the exact crystal structure (solid line) and the modified structure with the two nearest neighbor shells merged (dashed line). In (b), the original cohesive energy (solid line) is compared with the prediction of the modified pair potential in (a) (dashed line).

addition of a large number of quantities of nearly equal magnitude and alternating sign, a well-known source of numerical instability. The effect is greatest where the associated energy is greatest, *i. e.* in the first two neighbor shells. In the recursive approach, the same instability arises from an unstable feedback between the first and higher neighbor shells as the first neighbor distance is reduced from the cutoff.

In our set of crystals the instability is only significant for BC8, which has two nearly degenerate shells, $n_1 = 1$, $r = 2.31$ Å, $n_2 = 3$ and $s_2 = 1.03$. As demonstrated in Fig. 4.4 (a), the artificial oscillations can be completely removed by merging the first two shells, replacing them with a weighted average, $n_1 = 4$ and 2.37 Å. This has little physical effect since the bond lengths are not distorted much, and higher shells have exactly the correct structure. Indeed, if the energy of the unaltered crystal structure is computed with the pair potential stabilized by merging (the dashed line in Fig. 4.4(a)), then it is quite close to the original energy curve, especially near the minimum, as shown in Fig 4.4 (b). For BCT5 and β-tin, which also have their coordinations split across two shells like BC8, it turns out that merging is not required because the shell radii are sufficiently well-separated.

The example of BC8 shows that the CGE formula cannot be applied blindly to obtain reasonable results. It is somewhat surprising is that the two highly disparate pair potentials in Fig 4.4 (a) produce the quite similar cohesive energy curves in Fig 4.4 (b) for the BC8 structure. Clearly the pair potential before merging is not physically meaningful and is dominated by numerical instability. On the other hand, with a straight-forward modification that does not much disturb the physics, a reasonable potential is recovered that reproduces the energy curve quite well. No author has previously considered numerical instabilities during inversion of cohesive energy curves, presumably because because the problems are not severe as long as nearly-degenerate lattices are avoided (as they have been before this work). However, we shall see that in the case of many-body inversion issues of numerical instability simply cannot be

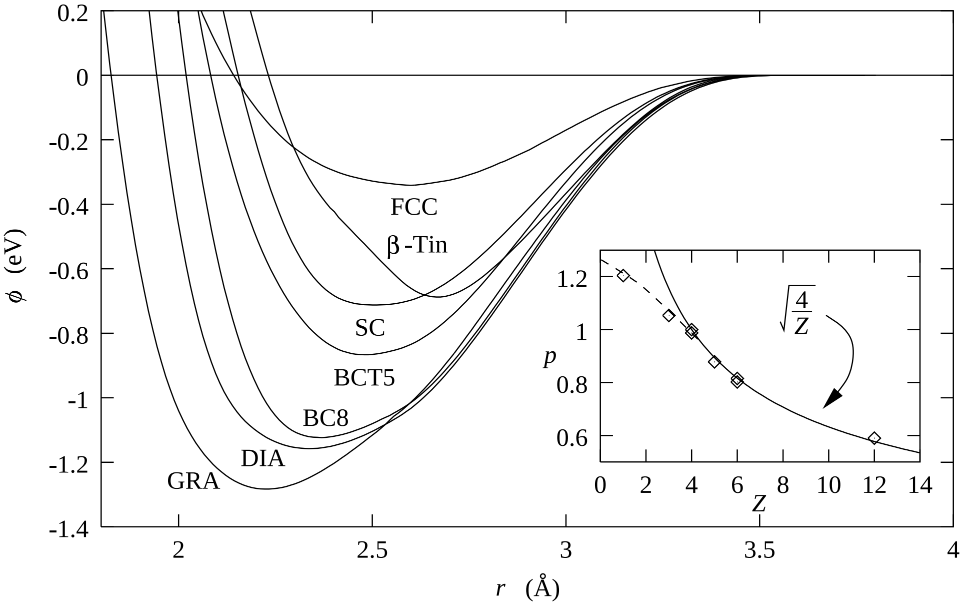


Figure 4.5: Inverted pair potentials (with cutoff) for seven silicon bulk phases. The inset shows the implied bond order p extracted from these curves (points) compared to $\sqrt{4/Z}$ (line). $p(1)$ reflects the Si_2 bond length and energy [19].

overlooked.

4.1.9 A Study of Pair Bonding in Silicon

Having analyzed the physical validity and numerical stability of pair potential inversion, let us apply our techniques to the crystal data set. The potentials shown in Fig. 4.5 are obtained by applying the cutoff to all of our energy curves and inverting. Once again, the large discrepancy between them is direct evidence for the well-known fact that the energetics of silicon cannot be described by a pair potential alone [17]. These results do suggest, however, that an environment-dependent pair potential can describe the energetics of ideal bulk phases reasonably well. There is a clear coordination dependence to the curves: bond lengths (positions of the minima) increase, and bond strengths (depths of the minima) decrease with increasing coordination.

This behavior can be understood within the bond-order formalism, which is justified on grounds of theoretical arguments [16, 17, 20, 67, 63, 62, 65, 66] as well as experience with empirical potentials [18, 37, 38, 44, 56]. In its simplest form, a bond order potential takes the form of an leading to an environment-dependent pair interaction [101, 17, 63],

$$\phi(r, Z) = \phi_R(r) + p(Z)\phi_A(r). \tag{4.13}$$

where $\phi_R(r)$ represents the short-range repulsion of atoms due to Pauli exclusion of their electrons, $\phi_A(r)$ represents the attractive force of bond formation and $p(Z)$ is the bond order, which modulates the strength of the attraction as a function of the atomic environment, measured by the coordination Z. The theoretical behavior of $p(Z)$ is as follows: The ideal coordination for Si is $Z_o = 4$, due to its valence. As an atom becomes increasingly overcoordinated ($Z > Z_o$), nearby bonds become more metallic, characterized by delocalized electrons. In terms of electronic structure, the local density of states for overcoordinated atoms can be reasonably well described by its scalar second moment. It is a well established result that the leading order behavior of the bond order is $p(Z) \sim Z^{-1/2}$ in the second moment approximation [101, 17, 20, 66, 63]. For $Z \leq Z_o$ on the other hand, a matrix second moment treatment predicts a roughly constant bond order (additive bond strengths) [64]. For small coordinations higher moments are needed to incorporate important features of band shape characteristic of covalent bonding, primarily the formation of a gap in the LDOS [17, 20, 64, 65]. Thus, the bond order should depart from the divergent $Z^{-1/2}$ behavior at lower coordinations with a shoulder at the ideal coordination of $Z = Z_o$ where the transition to metallic $Z^{-1/2}$ dependence begins.

If we could somehow extract a value of the bond order for each inverted potential, then we could make unprecedented, quantitative comparisons between the *ab initio* data and chemical bonding theory. One way to accomplish this is to assume a form for the repulsive interaction $\phi_R(r)$. In that case the bond-order term can be obtained from the *ab initio* data using $p(Z) = V_A(r_o)/V_A^{dia}(r_o)$, where $V_A(r,Z) = \phi(r,Z) - \phi_R(r)$ and r_o is the minimum of the inverted potential $\phi(r,Z)$. (Set $p(4) = 1$ for the diamond lattice.) The repulsive term is the weakest link in bond-order models, since its form must be assumed and then fit to empirical data with little theoretical guidance. Thus, it is reasonable to explore what happens with different choices. Here we see an advantage of working with multiple cohesive energy curves for the same material: we can objectively test the validity of the *functional form* of Eq. (4.13). If the form is physically accurate for a wide range of volumes and coordinations, then there should exist a choice of repulsive term that causes a collapse of the attractive terms, $\phi_A(r) = V_A(r,Z)/p(Z)$, obtained from different crystals. In the metallic regime ($Z > 4$), where the theoretical prediction is $p(Z) \approx (4/Z)^{1/2}$, this corresponds to checking the collapse of the functions $(Z/4)^{1/2}V_A(r)$. Consistent with the earlier use of the SW cutoff, the SW repulsive term ϕ_R^{SW} was tested and reasonable collapse of the attractive functions for the metallic phases was found. Although the general trend is insensitive to the choice of ϕ_R, the collapse is improved by assuming a stronger repulsive term. The bond order as a function of coordination is plotted in the inset of Fig. 4.5 for the case of $\phi_R = 2\phi_R^{SW}$. With this choice, the inverted bond order is in superb agreement with theory, $p(Z) \approx (Z/4)^{1/2}$ for $Z \geq 4$ and $1 < p(Z) < (Z/4)^{1/2}$ for $Z < 4$.

The empirical Pauling relation between bond length and bond order, $r \propto -\log p$ [108], which has been derived from chemical pseudopotential theory by Abell [16], has also been investigated for the inverted potentials. As shown in Fig. 4.6, the Pauling relation is satisfied fairly well by all the data, with the exception of the FCC data point. A closer look at the FCC inverted potential in Fig. 4.5, however, reveals that the minimum of the inverted potential has been artificially moved to a shorter distance by proximity to the cutoff, which was chosen to be distant from the diamond, but not the FCC, equilibrium bond length. Correcting for this fact places the FCC point close to the linear fit of the data.

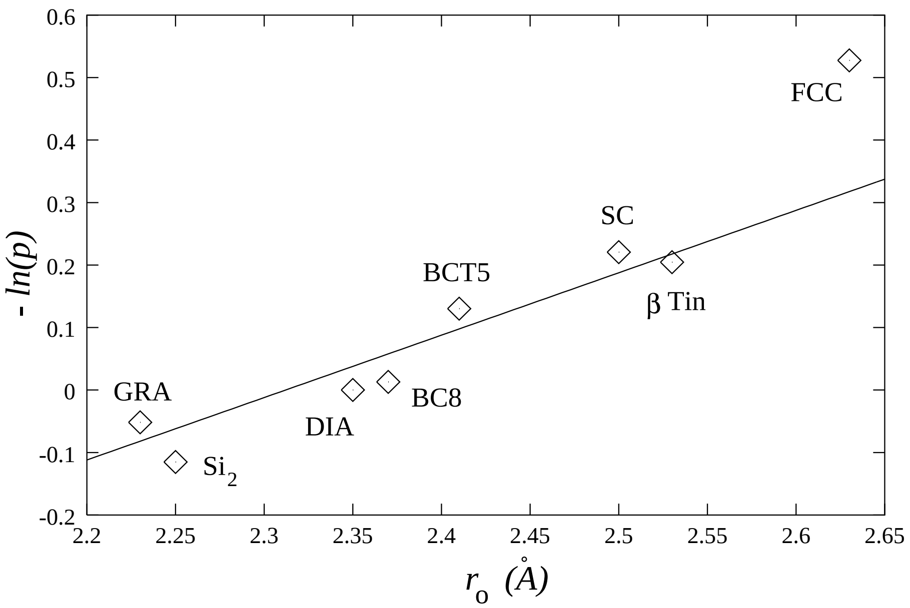


Figure 4.6: The negative logarithm of the bond order versus the radius of the minimum for our silicon inverted pair potentials. The linear fit indicates reasonable agreement with the Pauling relation between bond order and bond length.

4.1.10 Discussion

The results of the previous section provide *a posteriori* validation of the modifications we made to salvage meaningful results from inversion. In fact, this is the first inversion study to demonstrate any quantitative agreement with theory, which we have argued is a consequence of the physical insight afforded by the recursive formulation. The remarkable consistency between these results and bonding theory suggests that we might place enough confidence in the method to the reverse the logic: this is also the first direct, *ab initio* evidence for any material that the bond order form of the pair interaction is valid over a wide range of volumes and local bonding arrangements. Previous arguments supporting the bond order formalism have only come from crude, chemical trends in equilibrium bond lengths [16, 37, 45]. Thus, inversion has provided reliable first-principles information about the functional form of pair bonding interactions in silicon, something which has proven elusive when pursued with the ubiquitous fitting approach [19].

These results also have immediate implications for empirical potentials. The central conclusion is that the generic Tersoff format is much more realistic than the SW format for highly distorted configurations. This may seem to contradict the superiority of the Harrison format (which includes SW as a special case) for elastic properties demonstrated in the previous chapter. These findings are consistent, however, in light of Carlsson's argument that cluster potentials like SW can accurately fit narrow ranges of configurations while cluster functionals like Tersoff's provide a less accurate but physically acceptable fit to a much broader set of configurations (See Fig. 2 of Ref. [17]).

The inversion results also indicate that a coordination-dependent pair interaction can provide a fair description of high-symmetry crystal structures without requiring additional many-body interactions (unlike the Tersoff potentials which incorporate angular terms into the bond order). In particular, angular forces are only needed to stabilize these structures under symmetry-breaking distortions, primarily for small coordinations. In order to make a quantitative connection between Tersoff's functional form and our inverted *ab initio* data, angular contributions to the bond order must somehow be suppressed for ideal crystal structures, a point we shall revisit in Chapter 5.

4.2 Three-Body Cluster Potentials

4.2.1 A Three-Body Inversion Formula

Let us now generalize of the inversion procedure to the next order in the cluster expansion. Define the many-body component of the cohesive energy by subtracting off the pair contribution,

$$F_C(r) = E_C(r) - \sum_{p=1}^{\infty} n_p \phi(s_p r), \tag{4.14}$$

where C denotes the crystal structure. In the following derivation we must assume that $F_C(r)$ is known, i.e. that $\phi(r)$ can be determined, either theoretically or by inversion of $E_{C_0}(r)$ for some $C_0 \neq C$. The latter case is possible only if the angular dependence in C_0 makes the many-body terms vanish, but that is a reasonable case for tetrahedral solids[5]. As explained in Chapter 3, elastic constant relations suggest that many-body terms should vanish for the ideal diamond lattice.

The assumption of an environment-independent pair potential contradicts the strong evidence for the bond order coordination dependence given above, but let us make it anyway, just to see what we can learn. Any inconsistencies in the assumption should reveal themselves in the final results. It would be overly ambitious to try to invert

[5] Note that there is a minimum radius r_{min} of validity of the inversion, just below a/s_2^{dia} where second neighbors in the diamond lattice contribute to the energy. Since the angular function is usually nonzero for the (FCC) angles introduced by the second shell, it is no longer valid to use the inverted pair potential for the diamond lattice in constructing the many-body energy $F(r)$. For our choice of cutoff, however, this is not a major problem, since $r_{min} \approx 2.2$ is smaller than most covalent bond radii in silicon.

for environment-dependence and many-body interactions simultaneously. Although it is possible to include an assumed environment dependence in the pair (and three-body angular) interactions and still follow the procedure below, it is not in the spirit of the inversion method to make such assumptions, and we shall see that the simplest case of angular forces is already quite challenging.

More complicated cluster potentials and cluster functionals can be accommodated, but for simplicity let us consider separable three-body potentials of the form,

$$F[g,h](r) = \sum_i \sum_{j>i} g(R_i) g(R_j) h(\theta_{ij}), \tag{4.15}$$

where $\cos\theta_{ij} = \hat{R}_i \cdot \hat{R}_j$. This assumption is the starting point for the SW, Kaxiras-Pandey (KP) [42], and Biswas-Hamann (BH) [41] potentials, together with an environment-independent pair potential. Through inversion, we can study the validity of these common assumptions and test whether it is possible to derive a competitive potential of this form without any empirical inputs.

Suppose we are given an angular function $h(\theta)$. Our goal is then to find the three-body radial function $g[F,h](r)$ by inverting a cohesive energy curve. By performing many such inversions, we will see that the angular function can also be determined iteratively from the first guess. It is also possible to invert in the other order, as described in Appendix B: assume $g(r)$ and invert for $h[F,g](\theta)$. Although the latter approach is enticing, it has more restrictive problems with invertability, so we will proceed with the first approach.

Assuming then that we have $h(\theta)$, we can invert for $g[F,h](r)$ as follows: With A_p denoting the set of atoms in shell p, define,

$$\alpha_{pq} = \sum_{i \in A_p} \sum_{j \in A_q} h(\theta_{ij}), \tag{4.16}$$

where in the second sum, if $p = q$, then only $j > i$ should be considered to avoid double counting a triplet of atoms. With these definitions, the many-body contribution to the cohesive energy becomes,

$$F(r) = \sum_{p=1}^{\infty} \sum_{q=p}^{\infty} \alpha_{pq} g(s_p r) g(s_q r). \tag{4.17}$$

Separate the terms involving only $g(r)$,

$$\begin{aligned} F(r) &= \alpha_{11} g(r)^2 + \left[\sum_{p=2}^{\infty} \alpha_{1p} g(s_p r) \right] g(r) + \left[\sum_{p=2}^{\infty} \sum_{q=p}^{\infty} \alpha_{pq} g(s_p r) g(s_q r) \right] \\ &= \alpha_{11} g(r)^2 + \beta(r) g(r) + \gamma(r), \end{aligned} \tag{4.18}$$

where $\beta(r)$ and $\gamma(r)$ denote the corresponding terms in square brackets, giving

$$g(r) = \frac{-\beta(r) + \sqrt{\beta(r)^2 + 4\alpha_{11}(F(r) - \gamma(r))}}{2\alpha_{11}}. \tag{4.19}$$

The positive root is chosen in the quadratic formula, because the many-body energy should be positive [17]. As before, the idea is to view Eq.(4.19) as a recursion, since $g(r')$ appears in the expressions $\beta(r)$ and $\gamma(r)$. An explicit formula could be obtained by recursive substitution, but it involves a complicated set of nested square roots that is unwieldy to write down, even after the first recursive step. As in the pair potential case, it is much simpler to use the recursion directly in place of an explicit formula. The right hand side of Eq.(4.19) depends only on r' for $r' > r$, so we can solve for $g(r)$ in order of decreasing radius starting at the cutoff distance.

In principle, we can determine radial functions at any order of cluster expansion by applying this recursive approach to a family of cohesive energy curves, one for each radial function. For example, for a nonseparable three-body term, involving three bond lengths at once like the potential of Pearson *el. al.* [54], the recursion comes from solving a cubic equation, and for a four-body interaction, a quartic equation. Unfortunately, the numerical instabilities described below are magnified at higher orders, making such inversion intractable in practice. In the three-body case, however, we can obtain useful and physically reasonable results, thus for the first time incorporating both of the defining features of covalent solids, pair bonding and angular forces, into the inversion method. Before performing many-body inversion, however, much greater care must be taken than in the pair potential case, in order to overcome problems of invertability and numerical stability.

4.2.2 Existence of the Inverse

There are a several conditions that must be met in order for the inverse to exist. In the range $a/s_2 < r < a$, only first neighbors are inside the cutoff, so $\beta(r) = \gamma(r) = 0$. In this region, the recursion equation reduces to $F(r) = \alpha_{11} g(r)^2$. There are clearly two problematic cases: (*i*) $F(r)/\alpha_{11} < 0$ and (*ii*) $\alpha_{11} = 0$. Case (*i*) is a consequence of the assumed functional form: the three-body energy must have the same sign as the angular function for most neighbors in the first shell. Since theoretical approximations of quantum-mechanical models suggest that the three-body energy is positive, it is safest to choose non-negative angular functions, which is also consistent with theory [17, 66, 65]. This causes problems with our data for the graphitic lattice, because $F(r) < 0$ for all $r < a$. The non-invertability of the graphitic cohesive energy curve is a fundamental inconsistency of the assumption of an *environment-independent* cluster potential that we will address later, but it does not indicate a flaw in the inversion procedure.

Case (*ii*) is more subtle. A necessary (but not sufficient) condition for $\alpha_{11} \neq 0$ is that the first neighbor shell contain at least two atoms, n_p, forming at least one angle from the shortest bonds. In disordered structures, for which pair inversion might succeed, this condition may not be met, causing many-body inversion to fail. In our set of silicon crystals, there is one problematic case, BCT5, which only has one neighbor in the first shell at $r = 2.31$ Å, and four more in the second shell at $r = 2.43$ Å, for a total coordination of five. In order to avoid throwing this important structure out of the inversion set, it is necessary to merge the first two shells, keeping all angles and α_{ij} the same and replacing the first two shells by a single shell with $n_1 = 5$ and $r = 2.41$ Å, the weighted average of the original distances.

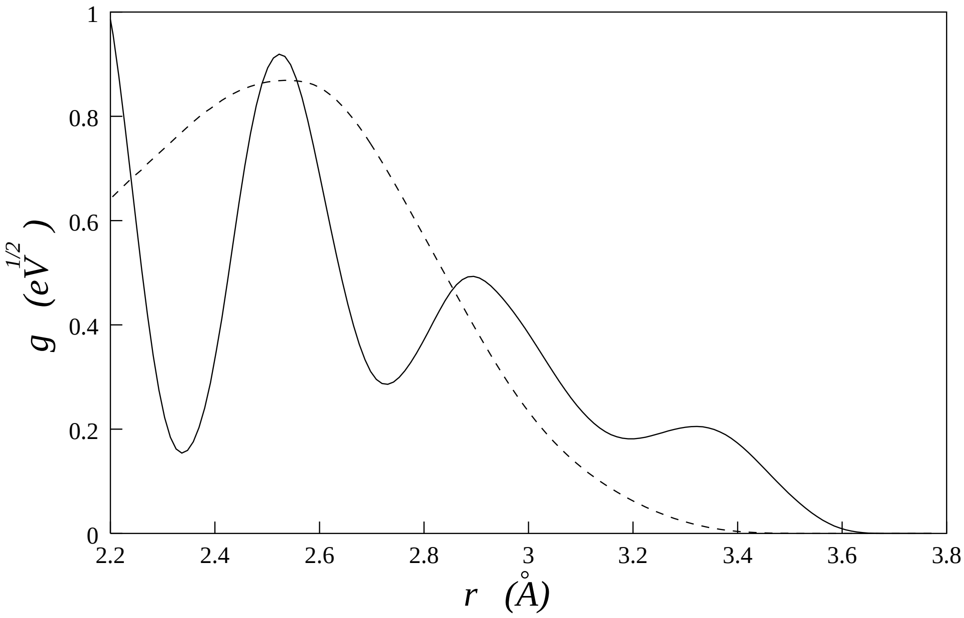


Figure 4.7: Numerical instability of three-body inversion. The inverted radial function for β-tin silicon, assuming the Stillinger-Weber angular dependence, is shown before (solid line) and after (dashed line) a cutoff function is applied to the many-body energy curve, showing how numerical instability can be controlled.

4.2.3 Numerical Stability

Even when the inverse exists, there can be artificial numerical instabilities. One general source of instability we have already encountered in the pair potential case: a second shell with comparable or greater occupation than the first shell with only slightly larger radius. In such cases, the nonlinear recursion generates unstable feedback that can result in large artificial oscillations. This is a major problem for BC8, but the shell merging used in the pair potential case removes the instability in the three-body case too. For β-tin, the separation of the first two shells is large enough that merging is once again not required.

Another important source of instability for all crystals is an overly abrupt change in $F(r)$ at the cutoff. This arises when taking the difference of two quantities (the pair energy determined by diamond lattice inversion and the total energy for the lattice in question) that might have different asymptotic behavior, resulting in a sudden, artificial rise in $F(r)$ near the cutoff. This problem can be solved by multiplying $F(r)$ itself by another cutoff function of the same form as used for $E(r)$. Fig. 4.7 shows the crucial effect of applying the cutoff function to $F(r)$ with the sensible choice $\sigma_F = 4\gamma_{SW}/\delta = 8.3804$ to give $g(r)$ exactly the SW asymptotic dependence. The oscillations in the bare, inverted $g(r)$ for β-tin are seen to be caused by an abrupt change in $F(r)$ at the cutoff. This effect is entirely artificial because any connection with the *ab initio* energy curve is suppressed where the energy is forced to zero by the pair potential cutoff function. Thus, no *ab initio* data is disturbed, and we can safely salvage reasonable behavior by smoothing the many-body energy near the cutoff.

Since the parameters of the cutoff function are arbitrary, we must understand their influences on the results. The cutoff range (which we take to be a_{SW} throughout this chapter) does not qualitatively change the results as long as it is smaller than the second neighbor distance in the equilibrium diamond lattice, which is reasonable for covalent, bond-bending forces. Similarly, the smoothing range does not have a major effect as long as it is short enough to avoid disturbing energies near equilibrium and long enough to gently enforce the cutoff. The choice $\delta = 1.2$ satisfies these requirements. The decay rate σ_F, however, has a subtle effect on the inverted radial functions, as illustrated in Fig. 4.8. If the decay is too slow, as it is with $\sigma_F = 4$, the change in $F(r)$ at the cutoff is too abrupt, and hence some of the unstable oscillations from Fig. 4.7 are not effectively suppressed. If the decay is too fast, then an artificial bump is created in $F(r)$ near $r = a - \delta$, where the smoothing begins. Just like the bump at $r = a$ when there is no cutoff function, this bump also causes unstable nonlinear oscillations. The instability is minimized in this case by $\sigma_F \approx 8$, which is fortuitously close to the SW

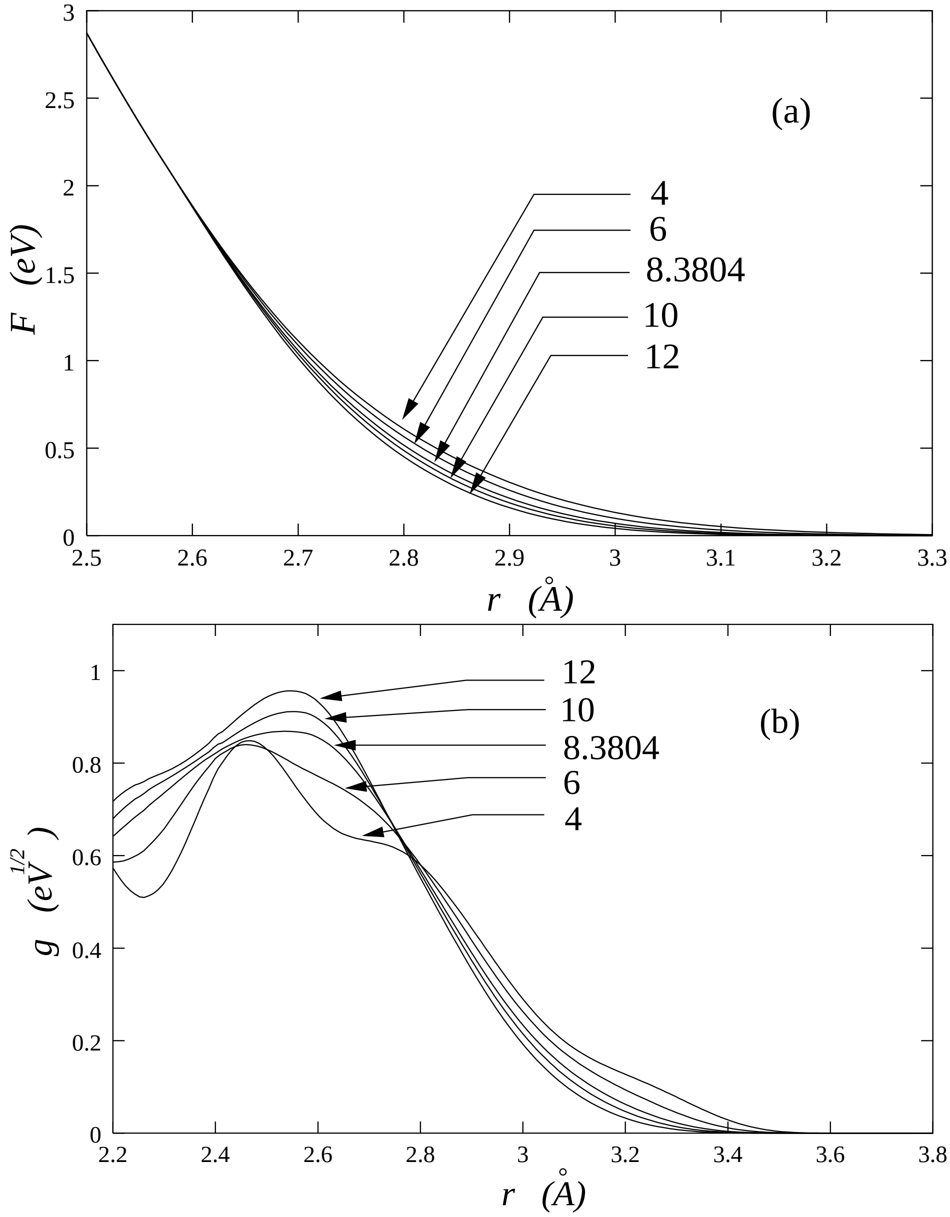


Figure 4.8: Dependence of three-body inversion on the many-body energy cutoff function. In (a), the many-body energy for β-tin silicon is shown for various values of the parameter σ_F controlling the decay of the cutoff function. In (b), the corresponding inverted radial functions are shown.

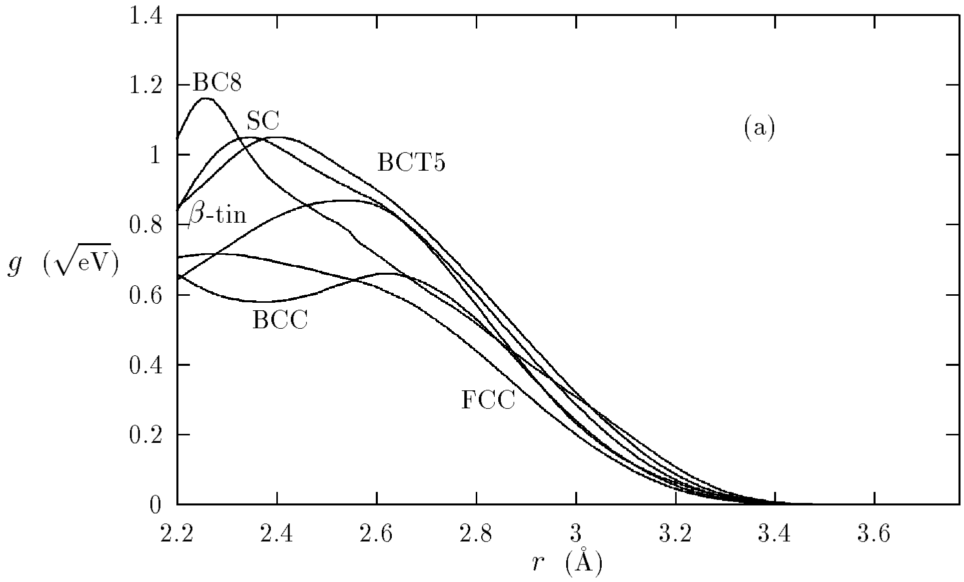
decay rate of 8.3804. It is remarkable how sensitive the three-body inversion procedure is to roughness in the many-body energy curve; the curves in Fig. 4.8 (a) look quite similar to the naked eye. In spite of this sensitivity, meaningful results can nevertheless be obtained for two reasons: (i) the decay can be systematically chosen as the one which introduces the least roughness, as measured by a quantity like, $\int_{a-\delta}^{a} (d^3F/dr^3)^2 dr$; ($ii$) the relative variation in the inverted radial functions introduced by the choice of decay is independent of the choice of angular function, thus not affecting the comparative analysis in the following section.

It is also worth noting that, in general, the smooth, short-range cutoff for many-body interactions is motivated by physical requirements, analogous the pair potential case. The angular dependence of cluster potentials is intended to describe bond-bending forces, primarily for sp^3 hybrid orbitals, in condensed phases. However, when the crystal is expanded so that the atoms are well-isolated, covalent bonding between hybrids is presumably replaced by a more spherically symmetric, metallic or van der Waals interaction [76]. Thus, the tails of cohesive energy curves are dominated by qualitatively different many-body interactions from condensed volumes, we would not expect an inversion procedure with long range to produce a physically meaningful three-body radial function.

4.2.4 A Study of Angular Forces in Silicon

To investigate angular forces in silicon from first principles, we perform many-body inversion for the following crystals, with the important modifications mentioned above: BC8, BCT5, β-tin, SC, BCC and FCC. As a first example, consider the SW angular dependence, $h(\theta) = s(\theta)^2 = (\cos\theta - \cos\theta_o)^2$, which vanishes at the tetrahedral angle $\theta_o = \cos^{-1}(-1/3) = 109.471°$. The inverted radial functions are shown in Fig. 4.9 (a). They bear some similarity to the fitted SW radial function, but there are important differences that may contain interesting physical information about angular forces. The radial functions tend to be peaked around the average distance of neighbors contributing to coordination. This property suggests that angular forces may be weakened if bonds are either stretched *or* compressed. There is also a coordination trend: angular forces are weaker for overcoordinated structures, which is consistent with the theoretical picture of the transition from covalent to metallic bonding. This kind of environment-dependence has not been included in any empirical potential (except for the model presented in the next chapter), and may lead to greater transferability between covalent and metallic phases.

These results depend on the *ad hoc* choice of the SW angular function, and thus we must next explore the effect of changing the angular dependence. In order to reliably use the diamond inverted pair potential, the angular function must vanish at the tetrahedral angle, which greatly narrows the class of angular functions we need consider. Luckily, this assumption is validated by the analysis of elastic properties presented in Chapter 3. (If not, we would not be able to obtain physically meaningful results from many-body inversion.) A simple variation on the SW angular function is to switch to an explicit dependence on the angle θ. Although the $\cos\theta$ dependence is suggested by quantum approximations [17, 65], it is interesting to investigate this possibility from first-principles. Fig. 4.9 (b) shows the inverted radial functions for the choice $h(\theta) = t(\theta)^2 = (\theta - \theta_o)^2$. Aside from the SC curve (and a slight overall change in scale), the inverted radial functions are quite similar in the two cases, indicating that these two angular functions have comparable consistency with the *ab initio* data for crystal phases. However, do not get the idea that the inverted radial functions are insensitive to the choice of angular function, as demonstrated in Fig. 4.10. For example, consider the square of the SW angular function, $s(\theta)^4$. Although this choice leads to a vanishing second shear modulus, there is no *a priori* reason to discount it for the large angular distortions present in our crystal structures. In that regime, no one really knows what the correct angular dependence is, or if even the concept of bond-bending is appropriate.



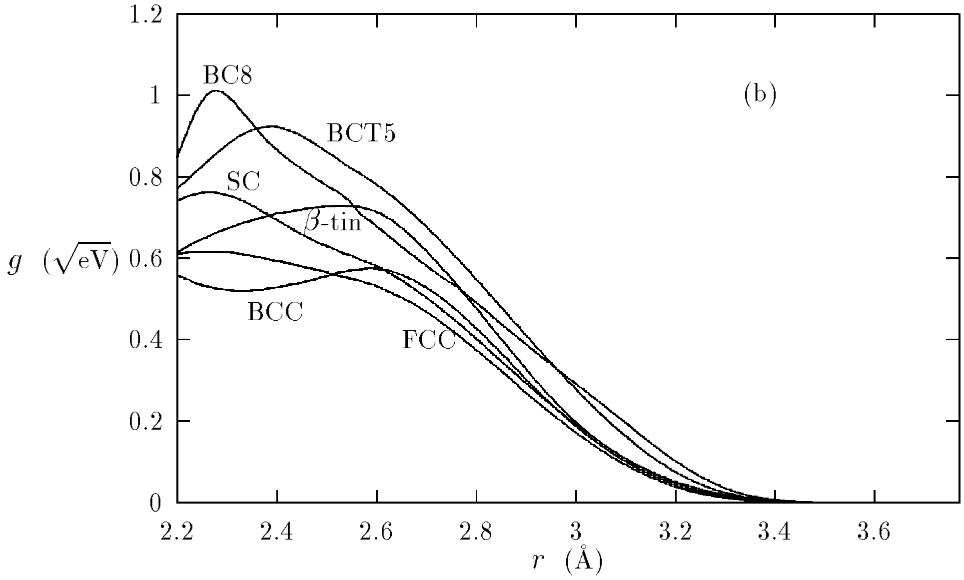


Figure 4.9: Inverted three body radial functions for silicon crystal structures, using (a) the SW angular function, $h(\theta) = (\cos\theta - \cos\theta_o)^2$, and (b) $h(\theta) = (\theta - \theta_o)^2$, where $\theta_o = \cos^{-1}(-1/3)$.

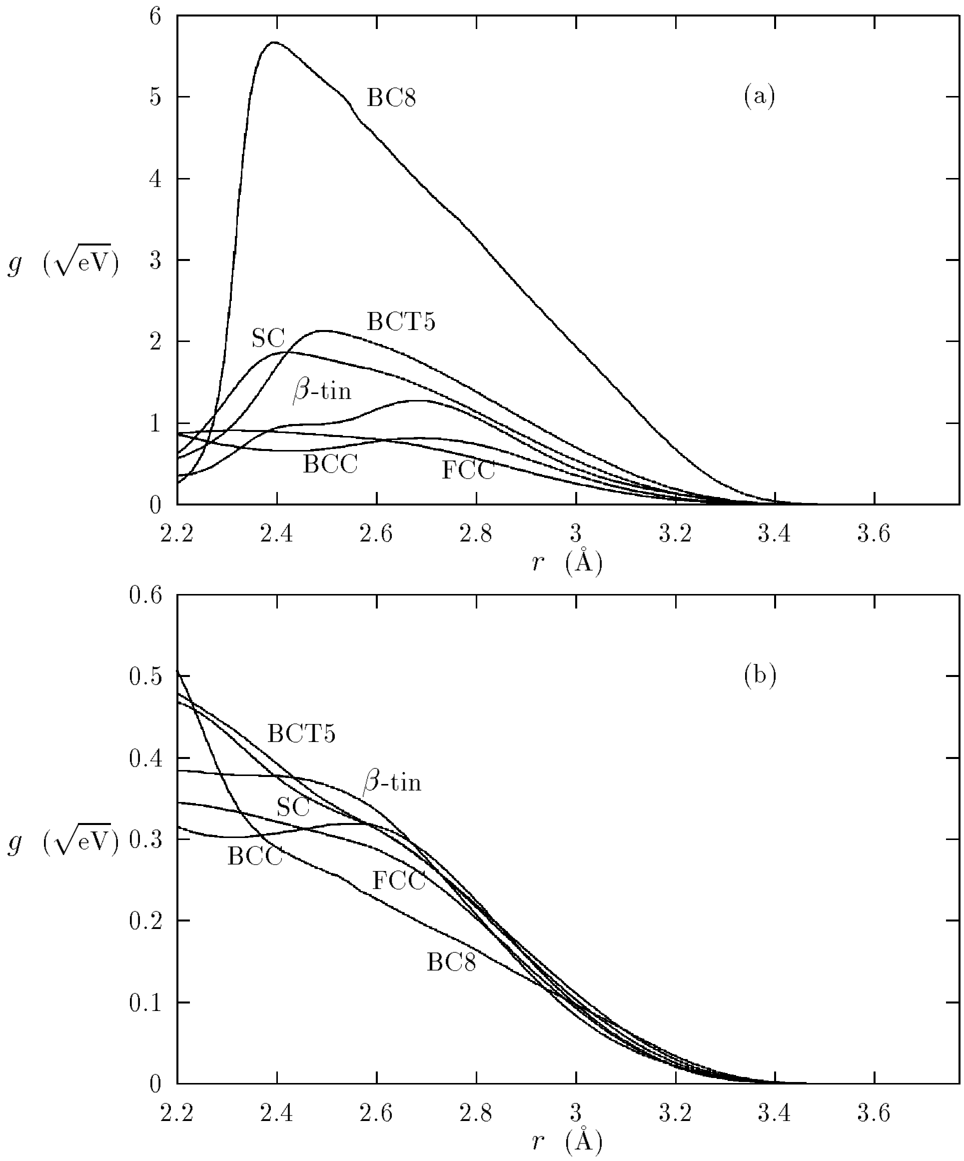


Figure 4.10: Inverted three body radial functions for silicon crystal structures, using (a) the square of the SW angular function and (b) the angular function of Kaxiras and Pandey.

As shown in Fig 4.10 (a), with this choice the inverted radial functions are much more varied than in the cases of Fig 4.9. The greatest anomaly is the BC8 curve, which is clearly a result of unphysical flatness of the angular function near the tetrahedral minimum, but the relative variation in the other curves is also considerably magnified compared with different choices of the angular function. On the other hand, the angular function of the potential of Kaxiras and Pandey [42], $h(\theta) = s(\theta)^2 - 0.894\ s(\theta)^4$, leads to a marked improvement in the collapse of the inverted radial functions.

The preceding analysis suggests a quantitative and parameter-free means to assess the quality of an angular function directly from the *ab initio* energy data: measure the variance of the inverted radial functions. If the angular function is physically correct (along with the underlying assumption of a three-body cluster potential), then the same radial function should result from every inversion, no matter what the input crystal structure. The relative spread of inverted radial functions can be measured with the quantity,

$$\Delta = \frac{\int_b^a \left(\frac{1}{N}\sum_{i=1}^N g_i(r)^2 - \overline{g}(r)^2\right) dr}{\int_b^a \overline{g}(r) dr}, \tag{4.20}$$

where $\overline{g}(r) = \frac{1}{N}\sum_{i=1}^N g_i(r)$ is the mean of the N inverted radial functions at distance r and $b = 2.2$ Åis the minimum radius of validity of the inversion (where second neighbors in the diamond lattice contribute to the pair energy). Division by the integrated mean deviation in the definition of Δ eliminates dependence on a multiplicative factor in $h(\theta)$, allowing for a fair comparison between functions with different shapes and sizes.

There are three interesting subsets of our data to consider in evaluating Δ. The first set, used in Δ_1, includes all invertible structures: β-tin, BC8, BCT5, SC, BCC and FCC. The second set, used in Δ_2, detects the effect of angles near tetrahedral by omitting BC8 from the first set, for a fair test of functions like $s(\theta)^4$ which are unreasonable for angles near tetrahedral but which may be appropriate for large angular distortions. The third set, used in Δ_3, includes only the three experimentally observed, low energy structures, β-tin, BC8, and BCT5.

Name	Angular Function	Δ_1	Δ_2	Δ_3
	$t(\theta)^2$	0.162	0.155	0.105
	$t(\theta)^4$	0.896	0.369	0.698
SW	$s(\theta)^2$	0.151	0.149	0.118
	$s(\theta)^4$	0.685	0.317	0.604
KP	$s(\theta)^2 + ks(\theta)^4$	0.113	0.075	0.125
INV	$s(\theta)^2 + c_1 s(\theta)^3 + c_2 s(\theta)^4$	0.108	0.088	0.073

Table 4.1: A quantitative comparison of candidate angular functions for silicon. The quantities Δ_i measure the ability of the angular functions to describe *ab initio* energy data for silicon bulk phases. $s(\theta) = \cos(\theta) - cos(\theta_0)$, $t(\theta) = \theta - \theta_0$, $\theta_0 = \cos^{-1}(-1/3)$, $k = -0.894$, $c_1 = -1.86$, and $c_2 = 1.423$. INV denotes the inverted angular function of Section 4.3.

Values of Δ_i for a variety of angular functions, including SW and KP, are displayed in Table 4.1. Note that the relative ordering of the angular functions is almost the same for any of the three Δ statistics, indicating that we might be getting a fair assessment of the relative physical validity of the angular functions. By comparing values of Δ_1 and Δ_2, we see that the leading term in a Taylor expansion of $h(\theta)$ about θ_o should be quadratic. Although that result is perhaps clear from elastic analysis, there is additional nontrivial information in Table 4.1. Expansions in $t(\theta)$ perform roughly as well as expansions in $s(\theta)$. This is somewhat surprising since the former have undesirable cusps at $\theta = 0$ and $\theta = \pi$ and have less theoretical motivation. The KP angular function, aside from being somewhat too flat near θ_o, collapses the radial functions quite well, but it has a deep, negative minimum at $\theta = 0$ which may produce spurious minimum energy structures. Finally, the data for the inverted angular function, described in the next section, shows that it is possible to improve on the other angular functions.

4.3 An *Ab Initio* Three-Body Potential for Silicon

4.3.1 Derivation

In order to complete the inversion procedure to obtain a parameter-free three-body cluster potential for silicon, optimal angular dependence can be extracted from the *ab initio* energy curves. This is accomplished by expanding the angular function in a series of SW-like terms, $h(\theta) = \sum_{i=0}^{3} c_i s(\theta)^{2+i}$, starting with the quadratic term for the reasons given above. We also set $c_0 = 1$, since an overall multiplicative factor has no effect. Additional terms are not included because the accuracy would be excessive considering the fairly small number of angles in our set of input structures[6]. The coefficients in the expansion are chosen to minimize the cost function $\Delta_3(c_i)$. The cost function is also augmented to penalize $h(\theta) < 0$ in order to avoid spurious minima at angles smaller than $\pi/3$, which are not present in our structures. In order to perform the minimization, simulated annealing[109] is employed because derivatives of the cost function are not available. During each annealing iteration the following steps are performed: (1) a small random change in the coefficient vector to select a candidate angular function h, (2) recursive inversion to obtain $g[h, F](r)$ for each crystal, (3) integration of these curves to evaluate the cost function, and (4) acceptance or rejectance of the random move with a Boltzmann probability factor, whose temperature is slowly reduced to drive the system toward a global minimum. The optimal angular dependence with $c_0 = 1, c_1 = -1.86, c_2 = 1.42$ and the corresponding radial function collapse are plotted in Fig. 4.11, and the final Δ_i values are given in Table 4.1.

A novel feature of the inverted angular function is its skew about the minimum to favor smaller angles, in contrast to most empirical potentials. This is consistent with

[6]To be specific, the angles between pairs of neighbors contributing to coordination in diamond, BC8, BCT5 and β-tin are: 86, 94, 99, 106, 109, 118, 148 and 149. FCC, BCC and SC widen the range of sampled angles, but they are omitted from the optimization because they are not low-energy bonding states in silicon.

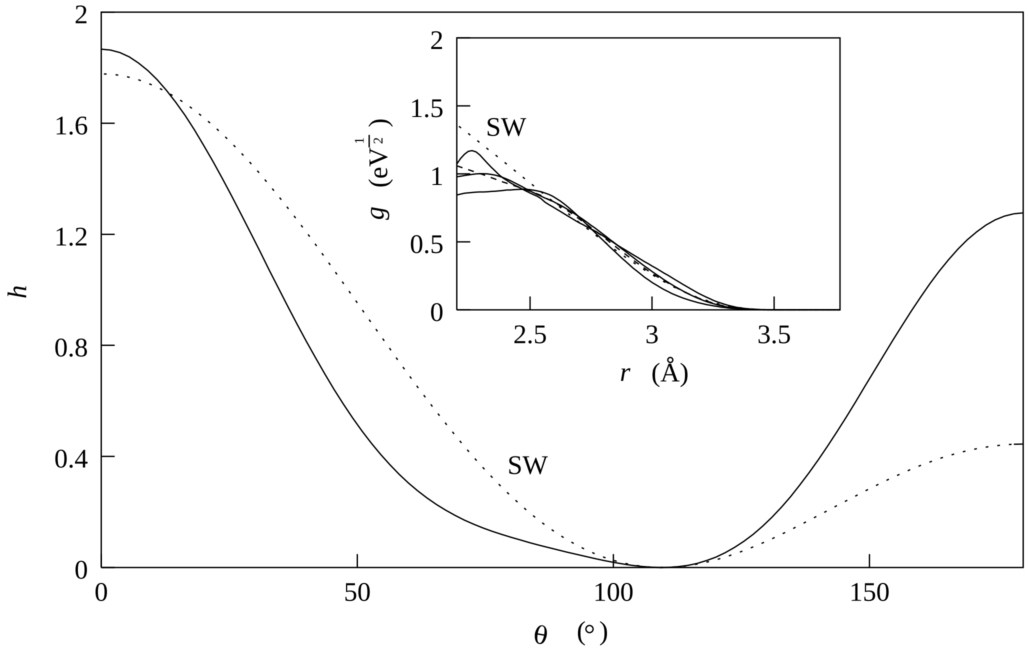


Figure 4.11: An inverted angular function for bulk silicon from the diamond, BC8, BCT5, and β-tin energy curves, compared with the Stillinger-Weber (SW) angular function. The inset shows the collapse of the inverted three-body radial functions with the average curve (dashed line) and the fitted SW radial function (dotted line).

the conclusion of a comparative study of potentials that angles smaller than $\pi/2$ tend to be overpenalized, which presumably leads to poor descriptions of surfaces, clusters, and certain defects [19]. The skewed angular function also raises the energy of overcoordinated metallic structures over covalent ones by penalizing large angles near π. It is typical to characterize metallic structures by the presence of small angles [110], but we observe that metallic structures also tend to have angles near π due to their cubic symmetry. Covalent bonds are actually characterized by angles in the intermediate range $\pi/2$ to $2\pi/3$.

At last, we have systematically arrived at an *ab initio*, parameter-free, three-body potential for bulk silicon. The pair interaction $\phi(r)$ is taken from diamond inversion, the "DIA" curve in Fig. 4.5. The angular function $h(\theta)$, shown in Fig. 4.11, comes from the optimized collapse of inverted three-body radial functions. By averaging the β-tin, BC8, and BCT5 radial functions, we obtain the radial function $g(r)$ of the inverted potential, the dashed line in the inset of Fig. 4.11. It is extended linearly to distances below the minimum radius of validity of many-body inversion, $r < 2.2$ Å. Both $\phi(r)$ and $g(r)$ are represented by 180 tabulated points in the range of 2.0 – 3.77118 Å. This is the first many-body potential to be inverted directly from *ab initio* cohesive energy calculations, with no empirical inputs.

4.3.2 Tests of the Inverted Potential

Elastic Constants

Although the potential is extracted from first principles data, it is not without assumptions, the most limiting being the form of a separable, three-body cluster potential. The physical validity of the underlying assumptions can be tested by checking the performance of the potential in various materials applications. We begin with the diamond structure. The lattice constant with the inverted potential, 5.397 Å, is a bit smaller than the experimental value of 5.43 Åpredicted by most empirical potentials, reflecting

	EXPT	LDA	INV	SW	T2
C_{11}	1.67		1.358	1.617	1.217
C_{12}	0.65		0.806	0.816	0.858
C_{44}	0.81		0.443	0.603	0.103
C_{44}^{o}		1.11	1.051	1.172	0.923
K	0.99		0.990	1.083	0.978
$C_{11} - C_{12}$	1.02		0.552	0.801	0.359
$C_{12} - C_{44}$	-0.16		0.363	0.213	0.755
ζ	0.62 - 0.75		0.702	0.629	0.83

Table 4.2: Elastic constants of the inverted potential (in MBar), compared with experiment (EXPT), *ab initio* (LDA), Stillinger-Weber (SW) and the second Tersoff potential (T2). The dimensionless Kleinman internal strain parameter ζ is also shown.

the slight underestimation of lattice constants by LDA. The elastic constants of the inverted potential are shown in Table 4.2. The bulk modulus is in perfect agreement with the *ab initio* value (by construction). The other elastic constants are not as good as with SW, but are better than with the second Tersoff potential (and with other popular models) [19]. For example, the inverted potential predicts C_{44} to be half the *ab initio value*, while SW and T2 predict three quarters and one eighth, respectively. As a result of the significant underestimation of C_{44}, none of these potentials can predict the negative Cauchy discrepancy. The Kleinman internal strain parameter, however, which expresses the effect of relaxation during shear strain, is very well described by the inverted potential.

Note that elastic constants are not included as *ab initio* input to the inverted potential, while most empirical models, with the notable exception of SW, have been explicitly fitted to elastic properties. As shown in Chapter 3, good elastic constants are characteristic of the functional form of the SW and inverted potentials. However, these results also demonstrate some transferability of this functional form because using unrelated properties are physical input seems to consistently produce reasonable elastic constants.

Crystal Stability and Phase Transitions

Next consider cohesive energy curves for the input structures computed with the potential, as shown in Fig. 4.12. This is a difficult test of the inverted potential, even though we are simply checking the input data, because the fitting problem is highly overdetermined. No existing empirical model provides a good description of all the important phases. The inverted potential (by construction) perfectly fits the diamond lattice data near the minimum. The diamond curve departs from the *ab initio* data at greatly expanded volumes due to the cutoff function, and also at greatly compressed volumes, due to breakdown of the assumption of negligible many-body energy. The BC8 data is also

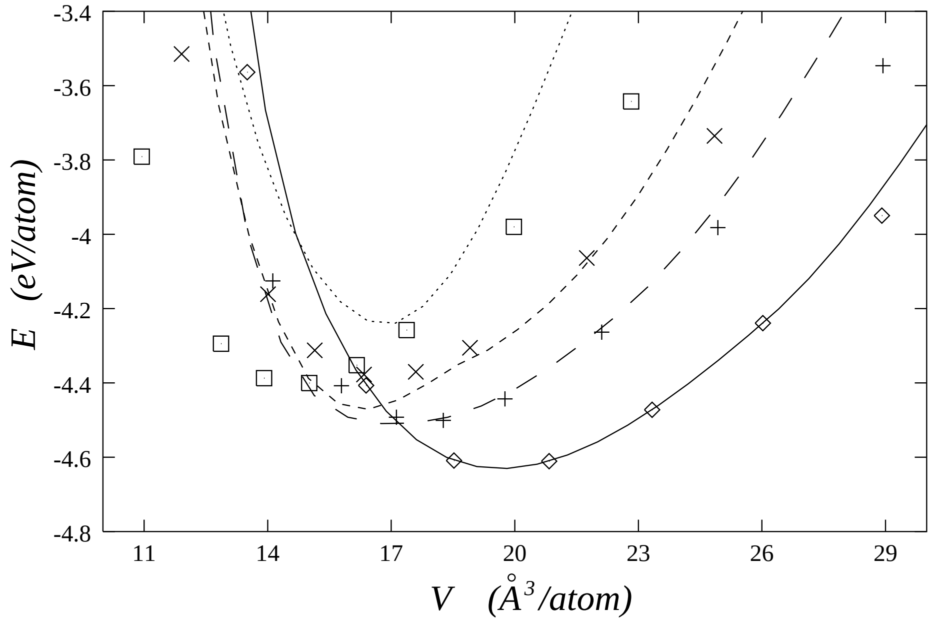


Figure 4.12: Cohesive energy curves computed with the inverted three-body potential for diamond (solid line), BC8 (long-dashed line), BCT5 (short-dashed line) and β-tin (dotted line) are compared with LDA data points for the same low energy structures, diamond (diamonds), BC8 (+), BCT5 ($\times$) and β-tin (squares).

well-reproduced across a wide range of volumes. The position of the minimum is somewhat low in volume, but not in energy. The BCT5 curve also reasonably close to LDA data points, but the β-tin curve is not so good. The latter's minimum is 0.2 eV high in energy and is also high in volume. The net result is that the first pressure-induced phase transition (from the tangent construction) predicted by the inverted potential is from diamond to BC8, with BCT5 being quite close[7]. The bulk modulus (curvature) of β-tin with the potential is 2.69 Mbar, more than twice the *ab initio* value of 1.18 Mbar.

Although these properties are not all favorable, the inverted potential performs as well as the best fitted models for crystal stability and phase transformations. The third Tersoff potential (T3) is the only known model to predict the diamond to β-tin transition, but it has many other problems [19]. The second Tersoff potential (T2), which performs considerably better than T3 overall, predicts a first transition to BC8, as does the SW potential and ours. However, ours is also only one of three potentials, along with SW and Dodson, known to give the correct ordering in energy of the experimentally accessible structures, diamond, BC8, and β-tin. The β-tin bulk modulus of the inverted potential, in spite of being large by a factor of two, is smaller than the values predicted by other models, with the exception of T3, which predicts 1.38 Mbar [19]. The most successful models overall, SW and T2, predict 4.43 and 3.40 Mbar, respectively.

Point Defects

Another important test of the transferability of potentials comes from point defects [19]. The formation energy of a vacancy with the inverted potential is equal to the binding energy, 4.63 eV, just like the SW potential. The *ab initio* value of 3.3 eV is smaller than the binding energy, indicating attractive second neighbor forces across the vacancy. Indeed, *ab initio* calculations show that the vacancy relaxes inward toward the

[7]Note that these results are not conclusive because internal relaxation was not performed in computing the cohesive energy curves. The *ab initio* relaxed structure was simply dilated. However, the qualitative results should not be greatly affected by relaxation.

defect along the $< 111 >$ direction, while the SW and inverted potentials do not relax at all[8]. The Tersoff potentials correctly predict the unrelaxed formation energy (2.83 eV for T2) to be smaller than the binding energy, but the Tersoff functional form incorrectly predicts outward relaxation, analogous to the (111) ideal surface reconstruction. The unrelaxed formation energies of the hexagonal and tetrahedral interstitials with the inverted potential are reasonably good, 4.9 eV and 2.2 eV, compared with LDA values of 4.3 and 3.7 eV, respectively. On the other hand, the SW values, 16.0 and 13.0 eV, are much too high. The reduced interstitial energies of the inverted potential versus SW are due to the greater tolerance for small angles ($\theta < \pi/2$) of the optimal angular function, as shown in Fig. 4.11.

An important activation energy in bulk silicon comes from the concerted exchange mechanism for self-diffusion [111, 42]. This complicated sequence of local configurations is traced out as two neighboring atoms exchange places in the diamond lattice by rotating about their common bond center. As shown in Fig. 4.13, the formation energy of the concerted exchange pathway computed with the inverted potential is fair agreement with the *ab initio* data, although there is a spurious, metastable minimum just above $60°$. The SW potential exhibits a less pronounced minimum, but SW overestimates the activation energy more than the inverted potential, which reproduces it quite well. Both SW and the inverted potential outperform the T2 potential, which predicts large, unphysical oscillations and does not even have a maximum at $90°$ [9].

[8] Both SW and the inverted potential relax inward to a more stable configuration if atoms are moved (*e. g.* due to thermal agitation) over a small energy barrier so that they are close enough for their dangling bonds to interact across the vacancy. There is no attraction, however, in the ideal configuration because the cutoff is smaller than the second neighbor distance.

[9] The best description of the concerted exchange path is provided by the fitted Environment-Dependent Interatomic Potential (EDIP) presented in Chapter 5. The EDIP curve is shown here to avoid repetition of this graph.

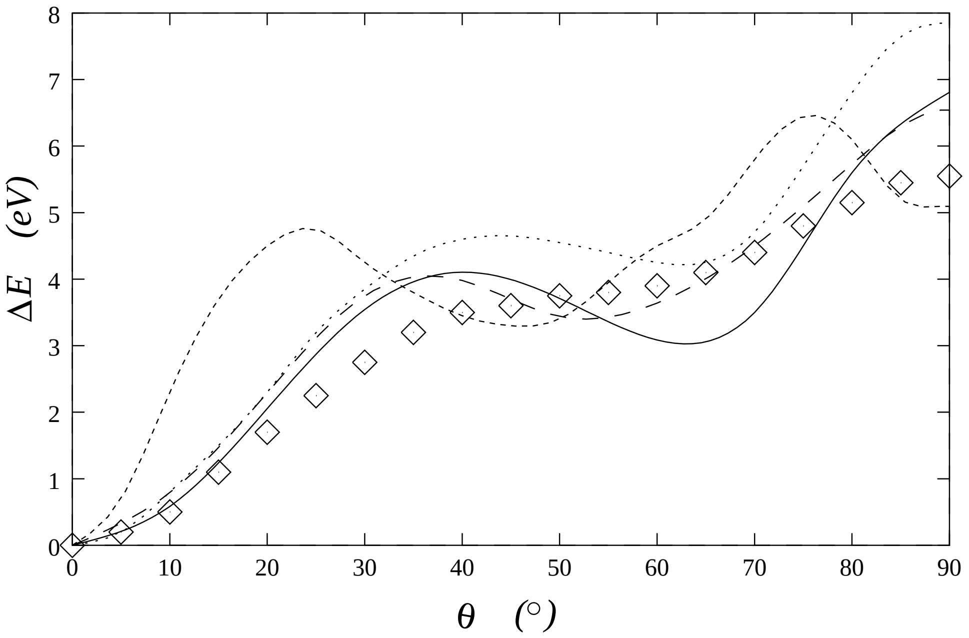


Figure 4.13: Energies of the concerted exchange mechanism for self-diffusion in silicon, as computed with the inverted potential (solid line), SW (dotted line) and T2 (short dashed line), and the EDIP potential of Chapter 5.2 (long dashed line), compared with the *ab initio* data (diamonds).

Surfaces

Finally, let us briefly consider the low energy (100) and (111) silicon surfaces. The tension of the ideal (100) 1×1 surface of 2.315 eV per unit cell is half the binding energy, due to the two broken bonds. Other potentials without environment dependence (like SW) make the same prediction, which consistent the LDA value of 2.5 eV [112][10]. The T2 potential strengthens the back bonds of the surface atoms, reducing the surface tension to 2.015 eV per unit cell. The change in surface tension for the symmetric 2×1 dimer reconstruction with the inverted potential is at least -0.971 eV per unit cell (not fully relaxed), compared with -0.899 eV and -1.258 eV for the SW and T2 potentials, respectively. The agreement with the *ab initio* value of -0.93 eV per unit cell is quite good for all the potentials because the reconstruction is a simple consequence of forming a surface bond from two adjacent dangling bonds [113]. A much more stringent test of the transferability of potentials to surfaces comes from the (111) surface. The SW and inverted potentials predict an ideal surface tension of 1.158 eV per 1×1 unit cell, which is one quarter the binding energy due to the single dangling bond per surface atom. The T2 potential again strengthens the three back bonds per surface atom for a lower surface tension of 0.707 eV per unit cell. The *ab initio* value of 1.56 eV per unit cell (which may again be an upper bound [19]) is considerably higher. More seriously, however, the inverted potential, like SW, does not predict stable adatom structures, which are essential ingredients of most (111) surface reconstructions, including the ground state 7×7 dimer-adatom-stacking fault reconstruction. For example, the 2×2 T_4 and H_3 adatom reconstructions are higher in energy than the ideal surface by 0.198 and 0.327 eV per unit cell, respectively, with the inverted potential. The SW values of 0.333 and 0.191 eV are equally bad. The T2 potential predicts stability with energy differences of -0.081 and -0.115 eV per cell, but these values are quite far from the *ab initio* values

[10] This is really only an upper bound since this early calculation was done with a rather small plane-wave energy cutoff of 4.3 Ry.

of -0.44 and -0.33 eV per cell, respectively, and are in the wrong order. Overall, the inverted potential is not well suited for surfaces, but it is not much worse than other empirical models. Since the input of the inverted potential is restricted to ideal bulk structures, we would not expect it to perform well for surfaces, so these results are not problematic.

In summary, the performance of the inverted potential for silicon bulk structures, elasticity, crystal phase transitions, point defects, concerted exchange and surface reconstructions appears to be comparable to that of the most popular empirical potentials (although the latter are much more thoroughly tested overall). It is remarkable that this performance for a wide range of noncrystalline defect structures has been achieved through the inversion of cohesive energy curves for ideal bulk phases with no additional empirical inputs. We may conclude that a great deal of information about bulk chemical bonding is contained in the cohesive energy curves of crystal phases.

4.3.3 Discussion

The closeness to first principles and relative simplicity of the inversion method make it an attractive alternative to the laborious and uncontrolled fitting approach, but it appears that the class of functional forms that can be used is rather limited by modern standards. We have seen that a three-body cluster potential which competes with SW and T2 can be derived through inversion alone. This would have been a breakthrough in 1984, but, with more than 30 potentials in the literature since then, a higher degree of sophistication and accuracy is currently required. However, just because inversion cannot be used as a black box to generate superior potentials does not mean it is without value. On the contrary, a great deal of useful qualitative and quantitative information can be gained from the analysis of inverted potentials, which can then be used to guide the selection of functional forms and their subsequent fitting. Inversion effectively removes some guesswork from the process of developing potentials and also increases our physical intuition about chemical bonding.

Our experience with three-body interactions contains a number of useful lessons. Although it it not always the case, inverted radial functions $g(r)$ tend to be strictly decreasing functions (like SW), especially when an overdetermined set of input structures is used. They also typically are fairly flat near the first neighbor distance, with a sharper decrease at large radii, indicating that angular forces stay strong even as covalent bonds are stretched and compressed by up to 10% or so. The rise of $g(r)$ at the cutoff also must be very gentle. Inverted angular functions $h(\theta)$ tend to penalize small angles ($\theta < \pi/2$) less than most existing models. Even when the angular function is adjusted to best collapse inverted radial functions, there is a clear coordination trend: the strength of angular forces decreases with increasing coordination, consistent with a transition from direct covalent bonding to more spherically symmetric metallic bonding. The inability of the inverted potential to describe this trend (constrained by its functional form) is evidenced by the predicted cohesive energy curves in Fig. 4.12. The covalent diamond and BC8 curves fit the *ab initio* data points quite well; the mixed covalent-metallic BCT5 curve departs somewhat from the *ab initio* data; and the metallic β-tin curve is quite far off, with excessively high energy and volume. A reasonable interpretation is that as coordination is increased the functional form cannot adapt to the changes in chemical bonding. The form of a three-body cluster potential can describe covalent four-fold coordination quite well, but has excessively strong angular forces for overcoordinated structures that artificially prevent the increase in density characteristic of metallic phases.

4.4 Conclusion

We have seen that with a number of innovations, stemming from the idea of recursion, inversion of cohesive energy curves can be raised from a theoretical curiosity to a systematic method of practical use in designing empirical potentials and understanding chemical bonding in covalent solids. We have explored for the first time important issues of physical validity and numerical stability. The transferability of inverted potentials for covalent solids has been improved by focusing only on condensed volumes and by considering multiple cohesive energy curves for different phases of the same material. As a proof-of-principle demonstration, we have derived a competitive three-body potential for silicon, a notoriously difficult case when pursued with empirical fitting schemes. Inversion has also revealed that environment-dependence is the key to improve upon current models. First-principles evidence has been given in support of the bond order form of the pair interaction for a wide range of ideal crystal structures at different volumes, and the softening of angular forces with increasing coordination is also suggested.

Together with the contents of Chapter 3, this body of results forms a reliable foundation upon which to build empirical potentials for bulk tetravalent solids. In general, we conclude that the functional form of atomic interactions should reduce exactly to appropriate cluster potentials in special bonding geometries, with environment dependence that interpolates smoothly between these special cases and captures general trends. Furthermore, we suggest that the most successful approach for designing superior potentials for silicon and related materials may be to use inversion to motivate an environment-dependent functional form, and then to use its quantitative predictions to guide the fitting process.