Interatomic Forces in Covalent Solids

A thesis presented

by

Martin Zdenek Bazant

to

The Department of Physics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Physics

Harvard University

Cambridge, Massachusetts

July 1997

Dedicated to my father,

Zdeněk Pavel Bažant

with love, respect and gratitude.

Abstract

One of the outstanding unsolved problems in the physics of materials is that of designing a transferable interatomic potential for covalently bonded solids, such as Si, Ge and C. In spite of intense efforts which have produced over thirty fitted potentials for the prototypical covalent solid, Si, realistic simulations are still problematic for important bulk phenomena such as plastic deformation, diffusion, crystallization and melting. In this thesis, innovative analytic techniques are used to extract concrete information regarding the functional form of interatomic potentials directly from *ab initio* energy calculations. By deriving elastic constant relations we study forces mediated by sp^3 and sp^2 hybrid covalent bonds, and by inversion of cohesive energy curves we explore the covalent to metallic transition and angular forces. This body of results provides a reliable foundation upon which to build empirical potentials and develop our intuition about chemical bonding. These theoretical predictions can be captured using a new functional form with only a few adjustable parameters called the Environment-Dependent Interatomic Potential (EDIP). Efforts to fit an EDIP for Si have already led to unprecedented transferability for bulk defects. Work in extending the model to disordered bulk phases (liquid and amorphous) is underway, and extensions to related materials should be possible. The speed of force evaluation with the new model is comparable to the most efficient existing potentials, making possible large-scale atomistic simulations of covalently bonded materials with heightened realism.

Acknowledgments

This work was generously supported by a Computational Science Graduate Fellowship from the Department of Energy along with grants from the Office of Naval Research and a teaching fellowship.

I have benefited from collaborations with a number of outstanding scientists whose influences permeate this thesis. Foremost, my advisor Efthimios Kaxiras has channeled my passion for mathematics into important applied problems with his insight and patience to let me explore the unknown. Under his guidance, I have learned not only a great deal of physics but also how to teach, work hard and communicate effectively. João Justo, Vasily Bulatov and Sidney Yip have contributed bright ideas and relentless effort to the development of the Environment-Dependent Interatomic Potential. I have also had the pleasure of working with the following people (on projects outside this thesis): Bruce Bayly (electrodeposition), Eric Heller (Lyapunov stability), Vasily Bulatov (dislocations) and Brond Larson (parallel computation). The faculty and students of Harvard University, particularly Adam Lupu-Sax, Noam Bernstein, Jan Fransaer, Normand Modine, Ellad Tadmor, Umesh Waghmare and Greg Smith, have also provided countless stimulating interactions.

Several dear friends have helped me persevere through difficult times, especially Adam Lupu-Sax, Douglas Reynolds and Christopher Ludford. That list also includes Sara Stevenson Bazant – our son Zdenek Stevenson Bazant is the greatest joy of my life. Born just before I began graduate school, “Steve” has contributed immensely to this work by taking me to the park day after day, spending the occasional night at my office and always reminding me of the important things. I also thank Nicole Weitzel for recently helping me through the final months of thesis writing.

Last but not least, my heartfelt appreciation goes out to my parents Zdeněk and Iva Bažant. Without their consistent emotional, intellectual and financial support, I simply could not have come this far.

Citations to Previously Published Work

Portions of Chapters 3–5 have appeared (or will soon appear) in the following published articles:

Chapters 3 and 5 : "Environment-Dependent Interatomic Potential for Bulk Silicon," M. Z. Bazant, E. Kaxiras and J. F. Justo, to appear in Phys. Rev. B (1997).

Chapter 4: "Derivation of Interatomic Potentials by Inversion of *Ab Initio* Cohesive Energy Curves," M. Z. Bazant and E. Kaxiras, in *Materials Theory, Simulations and Parallel Algorithms*, ed. by E. Kaxiras, J. Joannopoulos, P. Vashista, and R. Kalia, Materials Research Society Symposia Proceedings Vol. **408** (Materials Research Society, Pittsburgh, 1996), p. 79.

"Modeling of Covalent Bonding in Solids by Inversion of Cohehsive Energy Curves," M. Z. Bazant and E. Kaxiras, Phys. Rev. Lett. **77**, 4370 (1996).

Chapter 5: "Interatomic Potential for Condensed Phases and and Bulk Defects in Silicon," J. F. Justo, M. Z. Bazant, E. Kaxiras, V. V. Bulatov and S. Yip, Materials Research Society Proceedings, Spring Meeting Symposium E (Materials Research Society, Pittsburgh, 1997).

Contents

Title Page 1

Dedication 2

Abstract 3

Acknowledgments 4

Citations to Previously Published Work 5

Table of Contents 6

List of Figures 9

List of Tables 16

1 Introduction **20**

1.1 Interatomic Forces 20

1.2 Why Atoms Rather Than Electrons and Nuclei? 23

1.3 The Born-Oppenheimer Energy Surface 25

1.4 Cohesion in Covalent Solids 27

1.5 Scope and Outline of the Thesis 29

2 Models of Interatomic Forces in Covalent Solids **31**

2.1 Review of Empirical Potentials 32

2.2 Quantum-Mechanical Approximations 37

2.3 The Need for Theoretical Guidance 39

3 Elastic Constant Relations **41**

3.1 Taylor Expansion of the Cohesive Energy 43

3.2 Diamond sp^3 Hybrid Covalent Bonds 45

3.2.1 First Neighbor Interactions 45

3.2.2 Quantum-Mechanical Interpretation 51

3.2.3 Second Neighbor Interactions 55

3.3 Graphitic sp^2 Hybrid Covalent Bonds 59

3.3.1 *Ab Initio* Elastic Constants for Graphitic Silicon 59

3.3.2 First Neighbor Interactions 60

3.4 Comparison of sp^2 and sp^3 Bonds 62

3.5 Conclusion 63

4 Inversion of Cohesive Energy Curves **65**

4.1 Pair Potentials 66

4.1.1 The Carlsson-Gelatt-Ehrenreich Formula 66

4.1.2 The Chen-Möbius Theorem 68

4.1.3 Limitations of the CGE Formula 70

4.1.4 Recursive Inversion 71

4.1.5 *Ab Initio* Cohesive Energy Curves for Silicon Crystals 72

4.1.6 Physical Validity 75

4.1.7 An Essential Modification 78

4.1.8 Numerical Stability 79

4.1.9 A Study of Pair Bonding in Silicon 83

4.1.10 Discussion 87

4.2 Three-Body Cluster Potentials 88

4.2.1 A Three-Body Inversion Formula 88

4.2.2 Existence of the Inverse 91

4.2.3 Numerical Stability 92

4.2.4 A Study of Angular Forces in Silicon 95

4.3 An *Ab Initio* Three-Body Potential for Silicon 101

4.3.1 Derivation 101

4.3.2 Tests of the Inverted Potential 103

4.3.3 Discussion 111

4.4 Conclusion 112

5 The Environment-Dependent Interatomic Potential 114

5.1 Functional Form 116

5.1.1 Scalar Environment Description 116

5.1.2 Coordination-Dependent Chemical Bonding 118

5.1.3 Discussion 126

5.2 Fitting and Tests for Bulk Silicon 128

5.2.1 Fitting to Defect Structures 128

5.2.2 Tests for Bulk Properties and Defects 133

5.2.3 Cohesive Energy Curves 142

5.2.4 Discussion 146

6 Molecular Dynamics Simulation of Disordered Phases 147

6.1 Computational Methods 148

6.1.1 Scaling with System Size 149

6.1.2 Dynamics and Measurements 150

6.1.3 Efficient Force Computation 152

6.1.4 Benchmarks 157

6.2 The Liquid Phase 160

6.2.1 Crystal Melting 160

6.2.2 Liquid Structure 165

6.2.3 Discussion 173

6.3 Amorphous Phases 174
6.3.1 The Quenched Liquid 174
6.3.2 Another Amorphous Phase 178
6.4 Thermal Stability of Bulk Defects 182
6.5 Prospects for Increased Transferability 183

7 Conclusion **189**

Bibliography **195**

A The Geometry of Strained Diamond and Graphite **206**
A.1 Diamond Lattice 207
A.1.1 First Neighbors 207
A.1.2 Internal Relaxation 208
A.1.3 Second Neighbors 209
A.2 Graphitic Lattice 210

B Direct Inversion for Angular Forces **212**
B.1 Formulation of the Problem 213
B.2 The Constrained Case of an Even Angular Function 215
B.3 The General Case of a Skewed Angular Function 216
B.3.1 Volume-Scaled Reaction Pairs 217
B.3.2 Opening-Closing Reaction Pairs 218
B.4 Conclusion 219

C Recursion and the Möbius Inversion Formula **220**
C.1 A Recursive Approach to Möbius Inversion 221
C.2 A Combinatorial Expression for the Möbius Function 223

List of Figures

3.1 Quantum-mechanical effects during tetragonal yz (C_{44}) shear of the diamond lattice. The size of an atom (solid circle) suggests its position in the z direction out of the page. The equilibrium tetrahedron is shown in (a) with filled ellipses representing sp^3 hybrid orbitals. The strained state without relaxation is shown in (b) with rigid sp^3 hybrids, which we prove to be an accurate picture, at least for Si. The effects of internal relaxation in the x direction and rehybridization are shown in (c). . . . 53

3.2 *Ab initio* (LDA) data points (diamonds) for energy versus strain of stacked hexagonal planes with $c = 5.5$ Å. The dashed line is a fit of all the data with a sixth order polynomial. The solid line is a parabola fit to the linear elastic points (+). 60

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). 73

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. 76

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. 80

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). 81

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]. 83

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. 86

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. 92

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. 94

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 $\theta_o = \cos^{-1}(-1/3)$. 97

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. 98

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). 102

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). 106

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). 109

5.1 *Ab initio* values for the bond order in silicon as a function of coordination, obtained from the inversion of cohesive energy curves for the graphitic (GRA), cubic diamond (DIA), BC8, BCT5, SC, β-tin and BCC bulk structures and with additional points for the unrelaxed vacancy (VAC) and the dimer molecule (Si_2). For comparison the solid line shows the Gaussian $p(Z)$ obtained from fitting to defect structures. The dotted line shows the $1/\sqrt{Z}$ dependence, the theoretically predicted approximate behavior for $Z > 4$ (in the absence of angular forces). 120

5.2 Attractive pair interactions from inversion of *ab initio* cohesive energy curves for the structures in Fig. 5.1 using the bond order and repulsive pair potential of our model. The solid lines are for the covalent structures with coordinations 3 and 4, while the dotted lines are for the overcoordinated metallic structures. The reasonable collapse of the attractive pair potentials indicates the validity of the bond order functional form of the pair interaction across a wide range of volumes and crystal structures. . 121

5.3 The coordination dependence of the preferred bond angle $\theta_o(Z)$ (in degrees), which interpolates the theoretically motivated points for $Z = 2, 3, 4, 6$, indicated by diamonds. 124

5.4 The two-body interaction $V_2(r, Z)$ as a function of separation r for different coordinations: $Z = 3$ (dotted line), $Z = 4$ (small-dash line), and $Z = 6$ (large-dash line), compared with the SW (solid line) pair interaction.130

5.5 The three-body interaction $V_3(r, r, \cos\theta, Z)$ for a pair of bonds of fixed length $r = 2.35$ Å subtending an angle θ . The V_3 term is shown for several coordinations: $Z = 3$ (dotted line), $Z = 4$ (small-dash line), and $Z = 6$ (large-dash line), and compared with the SW (solid line) three-body interaction. 131

5.6 The function $f(r)$ that determines the contribution of each neighbor to the effective coordination Z. 133
5.7 Cross section of the $\{111\}$ glide set generalized stacking fault energy surface obtained from calculations using LDA (diamonds) the SW (dashed line) and EDIP (solid line) along the (a) $< 11\overline{2} >$ and (b) $< 1\overline{1}0 >$ directions. 139
5.8 Cohesive energy curves versus volume (a) and first neighbor distance (b) computed with EDIP for various silicon crystal structures. 143
5.9 Cohesive energy curves computed with EDIP, with the coordination number artificially fixed at the correct equilibrium value for each crystal. . . 144
5.10 Three-body radial functions $g(r)$ for silicon from (a) the fitting of EDIP and (b) the inversion of the cohesive energy curves. 145

6.1 Outline of an efficient algorithm to compute EDIP many-body environment-dependent forces. Indentation specifies the scope of a loop or conditional statement. Interpret "apply forces" throughout as "apply equal and opposite forces using Newton's third law". "Own" and "shared" refer to the SPMD parallel programming model. 156
6.2 Melting of a 1728-atom EDIP solid with a constant heat flux of 38.6 eV/atom·ns. The total (E), pair (V_2) and three-body (V_3) energies as a function of temperature are shown in (a), and the volume per atom versus temperature is shown in (b). 162
6.3 The melting transition (total energy versus temperature at constant pressure) under a constant heat flux for 1728-atom EDIP solids with (solid line) and without (dotted line) free surfaces. 164

6.4 Structure of the EDIP liquid. The pair correlation (solid line) is shown in (a) and compared with the SW liquid (dotted line), which is close to *ab initio* [151]. The bond angle distribution is shown in (b) for coordination neighbors with $r < 3.31$ Å (solid line) and also $r < 2.84$ Å (dashed line), and compared with the *ab initio* distribution for $r < 3.10$ Å (dotted line) and $r < 2.50$ Å (widely spaced dotted line) [151]. 167

6.5 A typical structure in the EDIP liquid, which may also appear in the *ab initio* liquid, possessing a mixture of covalent and metallic bonds. 169

6.6 The distribution of atomic coordinations in the *ab initio* (diamonds), EDIP (triangles) and SW (squares) liquids. The higher Z SW curve is for the true coordination cutoff 3.43 Å, and the other is for a shorter cutoff of 3.0 Å [134]. 170

6.7 Structure of the EDIP amorphous phase a-EDIP-I obtained from quenching the liquid. In (a), the EDIP (solid line) and *ab initio* (dotted line) [153] pair correlation functions are compared, along with peaks in $g(r)$ for a BCC crystal at the same density. In (b), the EDIP bond angle distributions for $r < 2.83$ Å (dotted line) and $r < 3.26$ Å (solid line) are shown, and compared with BCC angles. 176

6.8 Structure of the second EDIP amorphous phase a-EDIP-II obtained from annealing a SW amorphous sample (dotted lines). The pair correlation function is shown in (a), and the bond angle distributions for the first (solid line) and first two (dashed line) peaks of $g(r)$. 180

6.9 The pair correlation (a) and bond angle distribution (b) for amorphous phases of EDIP modified by using $H(x) = \lambda x^2$ with an extended three-body radial function (solid lines) and by simply setting $b = a$ (dashed lines), compared with the indirect SW structure (dotted lines). 184

6.10 The liquid pair correlation (a) and bond angle distribution (b) for EDIP modified by using $H(x) = \lambda x^2$ with an extended three-body radial function. 187

List of Tables

3.1 Comparison of elastic constants (in Mbar) for diamond cubic silicon computed with empirical models and the experimental (EXPT) or *ab initio* (LDA) values. The values for EXPT are from Simmons and Wang [78], for LDA from Nielsen and Martin [79] for tight-binding (TB) from Bernstein and Kaxiras [80] and for the empirical potentials Biswas-Haman (BH), Tersoff (T2, T3), Dodson (DOD) and Pearson-Takai-Halicioglu-Tiller (PTHT) from Balamane *et al* [19]. The Stillinger-Weber (SW) values are calculated with the analytic formulae of Cowley [71] and scaled to set the binding energy to 4.63 eV [19]. In the lower half of the table, the Born [9], Harrison [12] and new elastic constant relations are tested by calculating the ratios $\alpha_B \equiv 4C_{11}(C_{11}-C_{44})/(C_{11}+C_{12})^2$, $\alpha_H \equiv (7C_{11}+2C_{12})C_{44}/3(C_{11}+2C_{12})(C_{11}-C_{12})$ and $\alpha_{new} \equiv (4C_{11}+5C_{12})/9C^o_{44}$. . 48

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. 100

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. 104

5.1 Values of the parameters that define the latest version of EDIP for bulk silicon. 130

5.2 Elastic constants of the diamond phase of silicon in Mbar, from experiment (EXPT) [78] (and first principles for C^o_{44} [79]) compared with the predictions of empirical potentials EDIP and SW (from the formulae of Cowley [71]) and T3 [19], as well as a tight binding model (TB) [80]. The dimensionless Kleinman internal strain parameter is also compared with experiment ([82, 83, 84]. 134

5.3 Ideal formation energies E_f^{ideal} of point defects (in eV) and relaxation energies $\Delta E = E_f^{ideal} - E_f^{relaxed}$ with EDIP using a 54 atom unit cell, compared with *ab initio* (LDA) [111, 122, 123, 124], SW and T3 [19, 42] and tight-binding (TB) [80] results. 136

6.1 Timing analysis of our molecular dynamics program. The effects of thermodynamic phase (equilibrium liquid at $T = 2500$ K and solid at $T = 300$ K), system size (N particles) and small-scale parallelism ($P = 1$ or $P = 4$ processors) are demonstrated for systems in the microcanonical ensemble interacting via the SW potential on a Silicon Graphics RS-8000 Power Challenge. The total simulation time and force time are in μs per particle per time step, and percentages of the total time are given for force computation, neighbor list construction ($m = 100$ for solid, $m = 50$ for liquid, $\delta r = 0.03$ Å) and time integration (velocity Verlet scheme). . . . 158

6.2 Comparison of the speed of force computation with the Lennard-Jones (LJ), Stillinger-Weber (SW), Ismail-Kaxiras (IK) and EDIP potentials for typical solids and liquids (in μs per atom per time step on a Silicon Graphics RS-8000 processor). 158

6.3 Distribution of local coordinations (in %) for the *ab initio* (LDA) [151], SW ($r < 3.43$ Å) and EDIP ($r < 3.31$ Å) liquids at 2000 K. For EDIP, separate statistics are given for neighbors under the inner split first-neighbor peak of $g(r)$ (EDIP 1), and for SW we also show published data [134] (SW 1) presumably using a cutoff around 3.0 Å. The distribution of effective coordination numbers (EDIP Z) is also given, rounded to the nearest integer. 171

6.4 Coordination statistics for amorphous structures generated by *ab initio* (LDA) dynamics with quenching [153], the Wooten-Winer-Weaire algorithm (WWW) [155], the indirect SW method (SW) [133], and annealing of the indirect SW structure with EDIP, a-EDIP-II. 181

Chapter 1

Introduction

> If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words? I believe it is the *atomic hypothesis* (or the atomic *fact*, or whatever you wish to call it) that *all things are made of atoms – little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.* In that one sentence, you will see, there is an *enormous* amount of information about the world, if just a little imagination and thinking are applied.
>
> – Richard P. Feynman [1]

1.1 Interatomic Forces

The realization that matter is composed of tiny corpuscles called atoms is perhaps the greatest breakthrough in the history of science. The atomic hypothesis identifies the (usually) indivisible carriers of chemical identity and structure, which opens the possibility of predicting macroscopic materials phenomena from the microscopic level. Obviously, we could not understand chemical reactions like dissolution, catalysis and burning without talking about atoms because they are needed to identify the reacting substances, but the atomic hypothesis is also essential in cases not involving chemical changes. By thinking of matter as a collection of incompressible, indestructible atoms of finite size and mass that stick to each other, we can define physical concepts like heat (kinetic energy of atomic motion) and cohesion (potential energy of atomic arrangement). These ideas suffice for an intuitive picture of processes like sound propagation, evaporative cooling, melting, crystal growth, viscous fluid flow, solid deformation and fracture. Indeed, a central task of modern materials science is to understand macroscopic phenomena such as these in terms of the underlying atomic mechanisms.

The crucial property of atoms that determines such behavior is how they prefer to stick to each other; in other words, what are the interatomic forces? This question can be answered from first principles ("*ab initio*") by solving the quantum mechanical equations of motion for the atomic constituents, electrons and nuclei. While this is certainly the most reliable approach, there are two basic reasons to look for simpler descriptions that somehow capture the essential physics of the quantum mechanical treatment.

The first is that an *ab initio* solution of atomic motion is prohibitively expensive to calculate for more than a few hundred atoms, even on the fastest supercomputer. With simpler, classical models called *empirical interatomic potentials*, the same computers can perform simulations of millions of particles, making possible atomistic studies of incredibly complicated processes like melting, diffusion, sintering, amorphization, surface growth, radiation damage, cracking and plastic deformation in single crystals. The danger, of course, with such simulations is that in switching to the simple model, the essential physics has been lost, rendering the results meaningless. Certainly quantitative accuracy is reduced, but often even the qualitative behavior is incorrect.

The possible rewards of large-scale atomistic simulations, however, are so great that intense effort has been focussed recently on developing computationally efficient models of interatomic forces with increased realism. The goal is to make full use of high performance computers, which double in speed almost every year. If more realistic interatomic potentials can be designed, large-scale simulations may someday allow us to understand complex phenomena like the brittle to ductile transition, for example, from an atomistic level. That kind of knowledge would be extremely powerful in predicting materials properties and even engineering improved materials through computer experimentation.

While accurate, large-scale simulation is the usual motivation, there is another reason to develop simple models of interatomic forces that is rarely mentioned, namely to build our intuitive understanding of chemical bonding. Scientists like Cauchy, Poisson and Born were pondering the nature of interatomic forces long before the invention of the computer. Already then it was a nontrivial problem to explain experimental data, like cohesive energies and elastic constants, in terms of simple physical principles.

Today, the situation is much more challenging, because *ab initio* calculations, based on density functional theory [2] in the local density approximation (LDA) [3], have tremendously extended the body of accurate "experimental" data available that needs explanation [4]. Unfortunately, the output of every *ab initio* calculation is merely a number, the total energy of a particular atomic configuration[1]. The number is reliable, but we have little guidance in how to interpret it in terms of atomic interactions (or if such a thing is even possible within a simple framework).

In applied science, the primary goal is perhaps to predict physical properties with the maximum accuracy possible for the situation of interest. In more basic science, however, there is an inherent value in simple explanations, because a unified view of complex and seemingly disparate data can often be achieved.

[1]Actually, the meaningful output also includes the band structure and density distribution of the Kohn-Sham quasiparticles (which are believed to closely resemble the real, interacting electrons), but this information does not aid much in the quantitative understanding of interatomic forces.

> Although simplification is a sheer mathematical necessity, for many-body problems, there is also a more positive reason for it. What is it we really want from a theory? In the most interesting cases, what we are seeking is enlightenment, a general understanding of what is going on, a physical picture, something essentially qualitative that could be explained in relatively few words.... Simplification is an art rather like that of the cartoonist who captures the key features of a familiar face in a few deft strokes to make it instantly recognizable.
>
> – Sir Alan Cottrell [5]

One of our goals in this thesis is to represent a familiar covalent solid (Si) with a few "strokes" (or rather, potential energy functions) as deftly as we can, and see if it is still recognizable. Less colorfully, we aim to determine to what extent interatomic forces in covalent solids can be understood in simple terms and what degree of realism is possible with empirical potentials. We shall proceed by developing new methods of extracting classical interactions from *ab initio* data and by using that information, along with much blood, sweat and tears, to produce the best model we can. In the end, by seeing how well our model performs in a wide range of applications, we may learn something about the general strengths and limitations of interatomic potentials for covalent solids.

1.2 Why Atoms Rather Than Electrons and Nuclei?

Of course, we know that atoms are not really indivisible "little particles", so what do we mean when we talk about interatomic forces anyway? In the simple cases of noble elements or ionic solids, the physical picture given by Feynman above is quite accurate. Atoms (or ions) in these substances basically maintain their shape and chemical identity like spheres interacting via pairwise, radial attractions and repulsions. In the important cases of metals and semiconductors, however, the subatomic constituents, play crucial and vastly different roles in cohesion. One cannot begin to understand the subtleties of cohesion in these cases without considering electronic structure and its interplay with nuclear positions. Delocalization of electrons can also make the conceptual identity of an atom rather vague. For example, conduction electrons in a metal flow across macroscopic regions of space at enormous speeds averaging 10^{16} Å/s, typically undergoing collisions (strong interactions) with at least one out of every thousand atoms they pass. This means that in just one second, a typical metallic electron may be "shared" by over 10^{14} nuclei! Nevertheless, even though individual electrons and nuclei cannot be associated with each other, each nucleus is surrounded by a relatively static cloud of electron density, that may be conceptually divided among the nuclei to identify atoms.

In this regard, covalent solids, whose (semiconducting) electronic structure interpolates between the strongly localized (insulating) cases of noble or ionic solids and the delocalized (conducting) case of metals, are even more difficult to view from the atomic perspective. In these materials, valence electrons partially localize along "chemical bonds" with appreciable density concentrated in between bonded pairs of atoms. A valence electron in a bonding state is more or less equally shared between the two nuclei in the bond but may also resonate among a number of nearby bonding states. Thus, in a covalent solid the picture of well-defined atoms (resembling isolated atoms in a gas) sticking to each other at a preferred distance seems a bit strange. A covalent solid is more like a huge molecule, made up of around 10^{23} atoms.

In spite of these complications, however, the atomic picture of cohesion is justified in most condensed matter systems for one simple reason: nuclei are much heavier than electrons. The proton-electron mass ratio is 1836.15, and a Si^{28} nucleus is 5.157×10^4 times more massive than an electron. As a consequence, the fast and complex motions of the electrons are superimposed upon the (relatively) slow meanderings of the nuclei. Thus, if we identify each nucleus as the center of an atom, we can forget about explicitly keeping track of the electrons. Instead, we envision a "ball-and-spring model" of the material: the atoms (soft balls centered at nuclear positions) interact via an implicit force law (springs connecting the balls) determined by the electron density in the presence of the nuclei. Since core electrons stay tightly bound to the nuclei, it is more accurate to think of the atomic balls as representing ions consisting of the nuclei and their sheaths of core electrons, and the interatomic springs as the forces due to valence electron densities in the presence of the ionic pseudopotentials. In covalent solids, the springs can even assume a physical identity of their own, as chemical bonds, and we have then a simple language to describe atomic mechanisms in terms of the distorting, breaking and reforming of bonds. Since we do not need to solve explicitly for electronic structure in this approach, our task is to derive a classical interaction that somehow mimics the effect of the electrons on the nuclei. Although the problem is nontrivial, the reward for success is a tremendous conceptual and computational simplification.

1.3 The Born-Oppenheimer Energy Surface

An empirical interatomic potential is not just a toy model for qualitative understanding, akin to the Ising Hamiltonian for magnetic spin systems; instead, it can in principle provide a faithful *quantitative* reproduction of *ab initio* quantum mechanical predictions. This is a consequence of the adiabatic approximation, first applied to molecules by Born and Oppenheimer, which provides rigorous support for the arguments given above [6]. The adiabatic approximation justifies separation of the nuclear and electronic variables based on their disparate masses. The resulting errors in energy are smaller than the typical energy level spacings by a factor of order the mass ratio, less than 10^{-4} for most materials. Therefore, to a very good approximation, electrons move quantum-mechanically in a quasi-static external potential determined by the instantaneous positions of the nuclei, always in equilibrium due to their greater speeds. Conversely, the nuclei move in a force field determined by the time-averaged electron densities. In the context of molecules, the perturbative Born-Oppenheimer method or the variational method of Messiah may be used to derive precise *classical equations of motion for the nuclei* [6], and in the context of solids, the method of Car and Parinello [7] derives similar equations from self-consistent *ab initio* calculations of the instantaneous electronic ground state using density functional theory [2] in the local density approximation [3].

Whatever first principles method is used, the adiabatic approximation justifies the existence of the Born-Oppenheimer (BO) energy surface $E(\{\vec{R}_i\})$, which expresses the total energy of the system of electrons and nuclei as a function of the nuclear positions $\{\vec{R}_i\}$ alone. The force on atom i due to the presence of all the other atoms is simply $-\nabla_i E$, so *if* the BO surface were known, perhaps after being tabulated from many *ab initio* quantum-mechanical calculations, then atomistic dynamics could be performed using classical mechanics. Even though quantum mechanical equations of motion would not be integrated, the dynamics of the nuclei would be *exact* within the (very good) adiabatic approximation.

The difficulty is that *the BO surface is astronomically complicated*, except in special (very restrictive) cases. For example, in the simple case a 100 atom periodic lattice at finite temperature with up to 10% bond length distortions, in order to tabulate the total energy with a spatial resolution of 0.1% of the average bond length, we would need to perform a billion *ab initio* energy calculations. Now suppose we could somehow compile this data, it would still be a nontrivial task to design a data structure to store the massive table and an algorithm to access it efficiently. For more interesting situations involving larger systems with greater disorder, like the 1728 atom liquid simulations described in Chapter 6, it is clearly intractable to calculate, store and access the relevant regions of the BO surface, and no advantage over an *ab initio* quantum mechanical approach is achieved.

An obvious alternative is to design an empirical potential as follows: guess a simple functional form with adjustable parameters that allows efficient computation of forces, and fit it to a few carefully selected points on the *ab initio* BO surface. This approach is motivated by necessity, but there is no *a priori* guarantee that the potential is *transferable*, *i.e.* that it faithfully approximates regions of the BO surface to which it was not explicitly fit. Unfortunately, there is no small parameter, like $\hbar$ in semi-classical approximations, to bound errors, which may be unpredictably large or small in different cases. Very little theoretical guidance exists to select the correct form of an empirical potential. As a result, designing transferable potentials is a challenging and frustrating business, but, nevertheless, remarkable progress has been made for a wide range of materials [8].

1.4 Cohesion in Covalent Solids

The class of materials that has most resisted a transferable description by an empirical potential involves covalent bonding. In the prototypical case of silicon, over thirty potentials have been produced in recent years (reviewed in Chapter 2), but a satisfactory description has not been achieved, even for bulk material. To appreciate the subtleties involved in covalent solids, let us consider the simplest picture of interatomic forces, described above by Feynman, namely a pair potential in which atoms are attracted toward each other but resist being squeezed too close together. This kind of model, exemplified by the Lennard-Jones 6-12 potential for Van der Waals forces and the Coulomb electrostatic force law, captures the essential physics of noble, ionic, and, to a some extent, even metallic solids, but it is oversimplified in the case of covalent solids. For example, if we "apply a little imagination and thinking", we would predict qualitatively wrong crystal structures. The preference of atoms attracting via radial forces is to have as many neighbors as possible, since atoms simply want to be close to each other (up to a minimum distance). Thus, with pair potentials, crystal structure is mostly a matter of geometry, the close-packing of hard spheres in three-dimensions, which leads to lattices like face-centered cubic (FCC), hexagonal close-packed (HCP), body-centered cubic (BCC) and simple cubic (SC) with high density and coordination

(6–12). Covalent solids, however, crystallize in much more open structures like the diamond or graphitic lattices with low density and coordination (3–4), and usually *increase* their density upon melting.

Before the advent of quantum theory, the idea of pair potentials (radial forces) was advocated by influential scientists like Cauchy, and it was not until Born's seminal paper on diamond elastic constants in 1914 [9] that the need for non-radial forces to model covalent bonding was fully appreciated [10, 11]. He realized that additional forces are needed with explicit dependence on the *angles* subtended by the lines connecting atoms, not only to lower the energy of open lattices versus close-packed ones, but also to stabilize them against shear deformations. The model of Born was modified (for rotational invariance) and generalized by Harrison in 1956 [12]. Finally, in 1985 the conceptual framework of pair bonding and angular forces was extended to disordered structures with the potential of Stillinger and Weber (SW) [13], which has proven to be one of the most successful empirical models for covalent solids.

The original ideas of Born were motivated by elastic constant analysis, and thus are primarily relevant for small distortions of the diamond crystal structure. The SW potential illustrates, quite surprisingly, that the same concepts work fairly well for a broad range of configurations including crystal defects, liquid and amorphous states, but new ideas about the functional form of interatomic potentials are needed. The most obvious feature lacking in the SW model is *environment dependence*, or adaptation of the force law to changes in the local bonding environment. For example, liquid silicon is a metal with greater density and coordination than the solid, and metallic bonding is known in other materials to be described best by embedded-atom potentials [14], which usually have density-dependent bond strengths and no angular forces. It also seems unphysical that the SW model does not adapt to changes in covalent hybridization, for example, between the diamond and graphitic lattices. Experience with semi-empirical, quantum-mechanical (tight-binding) models has shown that transferability can be substantially increased by including environment-dependence [15], which suggests that the same may be true for classical, interatomic potentials.

Therefore, a crucial and ongoing theme in recent research is the environment dependence of interatomic potentials for covalent solids. Motivated by theoretical work [16, 17], environment dependence was first introduced by Tersoff in 1987 [18]. Since then numerous generalizations have appeared [19], and one version recently received theoretical justification from approximations of quantum theories [20, 21]. The next breakthrough in environment-dependence came with dangling bond vector of Chelikowsky *et. al.* [22], which is important for cases of broken lattice symmetry, like surfaces and clusters. In spite of these innovations, however, a significant improvement in transferability over the (much simpler) SW potential has not yet been achieved [19], which suggests that new ideas are needed to augment the Tersoff and Chelikowsky models. Unfortunately, the standard approach of fitting *ad hoc* functional forms has resulted in frustration, as increasingly complex and flexible functional forms have failed to yield substantial gains in transferability. Thus, new methods are needed to identify an improved functional form and to then guide the arduous fitting process.

1.5 Scope and Outline of the Thesis

This brings us to the questions we seek to answer in this work: Are there any indisputable facts about the functional form of an interatomic potential that can be deduced from *ab initio* calculations? Can potentials be derived directly from *ab initio* data, without fitting any adjustable parameters? How should environment dependence be included in the functional form? How well do theoretical results translate into accurate potentials, in practice? Is it really possible to attain a fully transferable description of a covalent material in all its important phases with a computationally efficient empirical potential, and if so, what methods might lead us to discover it?

We attempt to answer these important questions by developing new theoretical methods and applying them to the prototypical case of silicon. In recent years, Si has emerged as the representative covalent material due to its great technological importance and the vast amount of useful experimental and theoretical studies available to test new ideas. Our theoretical methods are equally applicable to related materials like Ge, C, and with minor extensions even alloys of these elements, but investigation of whether any of our specific results for Si carry over to these materials is beyond the scope of this thesis. Furthermore, a satisfactory description of bulk Si (crystal, defects, liquid and amorphous) has not yet been achieved, so here we shall focus on bulk interatomic forces, and postpone analysis of surfaces and molecules for subsequent work. Bulk Si already contains sufficiently complicated physics that we may use it to make progress toward answering our motivating questions.

We begin in Chapter 2 by comparing and contrasting existing empirical potentials for silicon, and mention some useful results from analytic approximations of quantum mechanical models. In Chapter 3, elastic constant relations are derived for various functional forms in the diamond and graphitic crystal structures to better understand interatomic forces mediated by hybrid covalent bonds. In Chapter 4, novel methods are developed to obtain many-body interatomic potentials directly from *ab initio* cohesive energy curves, which shed light on global changes in bonding across covalent and metallic structures. In Chapter 5, these theoretical results are incorporated into a new functional form called the "Environment-Dependent Interatomic Potential" (EDIP), which is fitted and tested for crystal phases and bulk defects. The computational efficiency and transferability of the fitted EDIP for disordered phases is studied in Chapter 6 using molecular dynamics techniques. Finally, Chapter 7 contains concluding remarks on our successes and failures and prospects for future research.

Chapter 2

Models of Interatomic Forces in Covalent Solids

Only quantum mechanics can account for the covalent bond.

– Andre Guinier [23]

In recent years, many empirical potentials for Si have been developed and applied to a number of different systems, and more recently compared to each other [19, 24]. Some of these models have been extended to other covalent materials, like Ge [136, 25], C [26, 27], F [28], S [29], SiGe [30, 31, 32], SiC [31], SiF [33], SiO_2 [34] and $GeSe_2$ [35], but by far the most testing of potentials has occured for Si, making it the ideal candidate for theoretical study into the fundamental issues of covalent bonding and representation by an empirical potential. Existing models differ in degree of sophistication, functional form, fitting strategy and range of interaction, and each can accurately model various special atomic configurations. Surfaces and small clusters are the most difficult to handle [19, 36], but even bulk material (crystalline and amorphous phases, solid defects and the liquid phase) has resisted a transferable description by a single potential. Realistic simulations of important bulk phenomena such as defect mobility, radiation damage, sintering, melting and crystallization are still problematic. In this chapter, we review existing potentials and approximations of quantum mechanical models in order to reach important conclusions about the desirable features of a successful interatomic potential.

2.1 Review of Empirical Potentials

The usual approach for deriving empirical potentials is to guess a functional form, motivated by physical intuition, and then to adjust parameters to fit *ab initio* total energy data for various atomic structures. A covalent material presents a difficult challenge because complex quantum-mechanical effects such as chemical bond formation and rupture, hybridization, metalization, charge transfer and bond bending must be described by an effective interaction between atoms in which the electronic degrees of freedom have somehow been "integrated out" [17]. In the case of Si, the abundance of potentials in the literature illustrates the difficulty of the problem and lack of specific theoretical guidance. In spite of the wide range of functional forms and fitting strategies, all proposed models possess comparable (and insufficient) overall accuracy [19]. It has proven almost impossible to attribute the successes or failures of a potential to specific features of its functional form. Nevertheless, much can be learned from past experience, and it is clear that a well-chosen functional form is more useful than elaborate fitting strategies.

To appreciate this point we compare and contrast some representative potentials for silicon. The pioneering potential of Stillinger and Weber (SW) has only eight parameters and was fitted to a few experimental properties of solid (cubic diamond) and liquid silicon [13]. The model takes the form of a third order cluster potential [17] in which the total energy of an atomic configuration $\{\vec{R}_{ij}\}$ is expressed as a linear combination of two- and three-body terms,

$$E = \sum_{ij} V_2(R_{ij}) + \sum_{ijk} V_3(\vec{R}_{ij}, \vec{R}_{ik}), \tag{2.1}$$

where $\vec{R}_{ij} = \vec{R}_j - \vec{R}_i$, $R_{ij} = |\vec{R}_{ij}|$ and we use the convention that multiple summation is over all permutations of distinct indices. The range of the SW potential is just short of the second neighbor distance in the ground state diamond lattice, so the pair interaction $V_2(r)$ has a deep well at the first neighbor distance to represent the restoring force against stretching sp^3 hybrid covalent bonds. The three-body interaction is expressed as a separable product of radial functions $g(r)$ and an angular function $h(\theta)$

$$V_3(\vec{r}_1, \vec{r}_2) = g(r_1)g(r_2)h(l_{12}), \tag{2.2}$$

where $l_{12} = \cos\theta_{12} = \vec{r}_1 \cdot \vec{r}_2/(r_1 r_2)$. The angular function, $h(l) = (l + 1/3)^2$, has a minimum of zero at the tetrahedral angle to represent the angular preference of sp^3 bonds, and the radial function $g(r)$ decreases with distance to reduce this effect when bonds are stretched. The SW three-body term captures the directed nature of covalent sp^3 bonds in a simple way that selects the diamond lattice over close-packed structures. Although the various terms lose their physical significance for distortions of the diamond lattice large enough to destroy sp^3 hybridization, the SW potential seems to give a reasonable description of many experimentally relevant states, such as point defects, certain surface structures, and the liquid and amorphous phases [19]. The SW potential continues to be a favorite choice in the literature, due in large part to its appealing simplicity and apparent physical content.

Another popular and innovative empirical model is the Tersoff potential, with three versions generally called T1 [18], T2 [37], and T3 [38]. The original version T1 has only six adjustable parameters, fitted to a small database of bulk polytypes. Subsequent versions involve seven more parameters to improve elastic properties. The Tersoff functional form is fundamentally different from the SW form in that the strength of individual bonds is affected by the presence of surrounding atoms. Using Carlsson's terminology, the Tersoff potential is a third order cluster functional [17] with the cluster sums appearing in nonlinear combinations. As suggested by theoretical arguments [39, 16, 20], the energy is the sum of a repulsive pair interaction $\phi_R(r)$ and an attractive interaction $p(\zeta)\phi_A(r)$ that depends on the local bonding environment, which is characterized by a scalar quantity ζ,

$$E = \sum_{ij} [\phi_R(R_{ij}) + p(\zeta_{ij})\phi_A(R_{ij})] \tag{2.3}$$

$$\zeta_{ij} = \sum_k V_3(\vec{R}_{ij}, \vec{R}_{ik}), \tag{2.4}$$

where the function $p(\zeta)$ represents the Pauling bond order. The three-body interaction has the form of Eq. (2.2) with the important difference that the angular function, although still positive, may not have a minimum at the tetrahedral angle. The T1, T2 and T3 angular functions are qualitatively different, possessing minima at 180^o, 90^o and 126.745^o, respectively. The original versions cannot describe the liquid and amorphous states[1]. The Tersoff format has greater theoretical justification away from the diamond lattice than SW, but the three fitted versions do not outperform the SW potential overall, perhaps due to their handling of angular forces [19]. Nevertheless, the Tersoff potential (or rather, family of potentials) is another example of a successful model for bulk properties with a physically motivated functional form and simple fitting strategy.

The majority of Si empirical potentials fall into either the generic SW [41, 42, 43] or Tersoff [44, 45, 46, 47, 48, 49] formats just described[2], but there are notable exceptions that provide further insight into successful approaches for designing potentials. First, a number of potentials possess functional forms that have either limited validity or no physical motivation at all, suggesting that fitting without theoretical guidance is not the optimal approach. The Valence Force Field and related potentials [50, 51, 52,

[1]This can be improved by changing the cutoff distance [40], but an uncontrolled change of the cutoff affects every other property of the potential, thus creating a new (and untested) model. It is sometimes said that "the Tersoff potential" can describe many properties (elastic constants, solid defects, surfaces, liquid, amorphous, clusters, ...), when in fact it been necessary to refit and modify the model for each circumstance, always at the cost of other desirable properties. The original SW potential, on the other hand, has provided reasonable transferability without any modifications.

[2]Most potentials for other covalent solids also assume either the SW [136, 25, 28, 29, 33] or Tersoff [26, 31, 32] functional form.

53] (of which there are over 40 in the literature [52]) involve scalar products of the vectors connecting atomic positions, an approximation that is strictly valid only for small departures from equilibrium. Thus, extending these models to highly distorted bonding environments undermines their theoretical basis. The potential of Pearson *et. al.* [54], as the authors emphasize, is not physically motivated, but rather results from an exercise in fitting. Their use of Lennard-Jones two-body terms and Axilrod-Teller three-body terms, characteristic of Van der Waals forces, has no justification for covalent materials. The potential of Mistriotis, Flytzanis and Farantos (MFF) [55] is an interesting attempt to include four-body interactions. Although the importance of four-body terms is certainly worth exploring, the inclusion of a four-body term in a linear cluster expansion is not unique, and theoretical analysis tends to favor nonlinear functionals [17, 16, 20].

A natural strategy to improve on the SW and Tersoff models is to replace simple functional forms with more flexible ones and complement them with more elaborate fitting schemes. The Bolding and Andersen (BA) potential [56] generalizes the Tersoff format with up to five-body interactions and over 30 adjustable parameters fit to an unusually wide range of structures. Although it has not been thoroughly tested, the BA potential appears to describe simultaneously bulk phases, defects, surfaces and small clusters, a claim that no other potential can make [19]. However, its complexity makes it difficult to interpret physically, and since a large fitting database was used, it is unclear whether the potential can reliably describe structures to which it was not explicitly fit. In this vein, the spline-fitted potentials of the Force Matching Method [57, 58] represent the opposite extreme of the SW and Tersoff approaches: physical motivation is bypassed in favor of elaborate fitting. These potentials involve complex combinations of cubic splines, which have effectively hundreds of adjustable parameters, and the strategy of matching forces on all atoms in various defect structures is the most elaborate attempted thus far. Although the method may be worth pursuing as an alternative, it has not yet produced competitive potentials [59]. Moreover, even if a reliable potential could result from such fitting strategies, it would make it hard to interpret the results of atomistic simulations in terms of simple principles of chemical bonding. Such interpretation is essential if any physical insight is to be gained from computer simulations.

In spite of relentless efforts, no potential has demonstrated a transferable description of silicon in all its forms [19] leading us to another important conclusion: it may be too ambitious to attempt a simultaneous fit of all of the important atomic structures (bulk crystalline, amorphous and liquid phases, surfaces, and clusters) since qualitatively different aspects of bonding are at work in different types of structures. Theory and general experience suggest that the main ingredient needed to differentiate between surface and bulk bonding preferences is a more sophisticated description of the local atomic environment. A notable example in this respect is the innovative Thermodynamic Interatomic Force Field (TIFF) potential of Chelikowsky *et. al.* [22], which includes a quantity called the "dangling bond vector" that is a weighted average of the vectors pointing to the neighbors of an atom,

$$\vec{D}_i = -\sum_{j\neq i} \vec{R}_{ij} f(R_{ij}), \tag{2.5}$$

where $f(r)$ is a cutoff function. For symmetric configurations characteristic of the ideal (or slightly distorted) bulk material, the dangling bond vector vanishes (or is exceedingly small). Conversely, a nonzero value of the dangling bond vector indicates an asymmetric distribution of neighbors. While the dangling bond vector description appears to be very useful for undercoordinated structures like surfaces and small clusters, in this thesis our focus is on bulk material and thus we only consider simpler, scalar environment descriptions. We shall see in the next chapter that the dangling bond vector also improves accuracy in the bulk, but it is not the leading order environment variable. Our review of the current state-of-the-art of empirical potentials shows that a goal of fundamental importance is to obtain the best possible description of condensed phases and defects with a simple, theoretically justified functional form.

2.2 Quantum-Mechanical Approximations

An alternative to fitting guessed functional forms is to derive potentials by systematic approximation of quantum-mechanical models. So far, this approach has failed to produce superior potentials, but important connections between electronic structure and effective interatomic potentials have been revealed. Although attempts are being made to directly approximate Density Functional Theory [60], the most useful contributions involve approximating various Tight Binding (TB) models, which can themselves be derived as approximations of first principles theories [61]. These methods are based on low order moment approximations of the TB local density of states (LDOS), which is used to express the average band energy as the sum of occupied bonding states [63, 17, 62, 20, 67, 64, 65, 66, 68]. Pettifor has derived a many-body potential, similar in form to the Tersoff potential, by approximation of the TB bond order [20]. More recently, an angular dependence remarkably close to the T3 angular function has been derived for σ bonding from the lowest order two-site term in the Bond Order Potential (BOP) expansion [21, 67], but the analytically derived function has a flat minimum at 120^o and thus differs qualitatively with the T1 and T2 potentials (the latter being the most successful version overall). A simple physical principle explains these results: a σ bond is most weakened (desaturated) by the presence of an another atom when the resulting angle is small ($\theta < 100^o$) because in such cases the atom lies near the bond axis, thus interfering with the σ orbital where it is most concentrated. Working within the same framework of the TB LDOS, Carlsson and coworkers have derived potentials with the Generalized Embedded Atom Method [64, 65, 66]. Harrison has arrived at a similar model by expanding the average band energy in the ratio of the width of the bonding band to the bond-antibond splitting, the relevant small parameter in semiconductors [68]. These potentials resemble the SW potential in its description of angular forces with an additive three-body term, particularly for small distortions of the diamond lattice. The transition to metallic behavior in overcoordinated structures involves interbond interactions similar to the Tersoff and embedded atom potentials.

Many-body potentials can be derived from quantum-mechanical models if we restrict our attention to important small sets of configurations. Using a basis of sp^3 hybrid orbitals in a TB model, Carlsson *et. al.* [17, 64] have shown that a generalization of the SW format, in which Eq. (2.2) is replaced by a form similar to that used by Biswas and Hamann (BH) [41],

$$V_3(\vec{r}_1, \vec{r}_2) = \sum_{m=0}^{2} g_m(r_1) g_m(r_2)\, l_{12}^m, \tag{2.6}$$

is valid in the vicinity of the equilibrium diamond lattice. In general, the fourth moment controls the essential band gap of a semiconductor, implying four-body interactions, but the separable, three-body SW/BH terms are a consequence of the open topology of the diamond lattice: the only four-atom hopping circuit between first neighbors is the self-retracing path $i \to j \to i \to k \to i$ [17].

We can make analogous arguments for the graphitic lattice to draw conclusions about sp^2 hybrid bonds. Ignoring the weak, long-range interaction between hexagonal planes, we can assume a TB basis of sp^2 hybrid orbitals and follow Carlsson's derivation. Because the self-retracing path is also the only first neighbor hopping circuit in a graphitic plane, a cluster expansion with the generic BH three-body interaction is also valid for hexagonal configurations, with the functions in Eqs. (2.1) and (2.6) differing from their diamond sp^3 counterparts, as described below. These calculations also suggest that a locally valid cluster expansion should acquire strong environment dependence for large distortions from the reference configuration [17].

In summary, these studies provide theoretical evidence that the linear three-body SW/BH format is appropriate near equilibrium structures, while the nonlinear many-body Tersoff format describes general trends across different bulk structures. For the asymmetric configurations found in surfaces and small clusters, these theories also suggest that a more complicated environment dependence than Tersoff's is needed, like the dangling bond vector of the TIFF potential [20, 64]. Direct approximation of quantum models can provide insight into the origins of interatomic forces, but apparently cannot produce improved potentials. The reason may be that the long chain of approximations connecting first principles and empirical theories is uncontrolled, in the sense that there is no small parameter which can provide an asymptotic bound for the neglected terms for a wide range of configurations. The expansion parameters in these studies are the dimensionless ratios of high to low order moments of the TB LDOS, which may not be small, especially for defect structures with states in the band gap or for metallic states like the liquid, the β-tin crystal structure, and certain surfaces. Low order moments capture general trends in energy, but cannot be expected to maintain quantitative accuracy.

2.3 The Need for Theoretical Guidance

A wide range of ideas about interatomic forces have been advanced and tested. Some are suggested by approximations of quantum theories or by empirical trends in chemical data, but most merely reflect physical intuition. Improving upon current models remains a supremely frustrating proposition. Increased flexibility and sophistication in fitting does not seem to help; the most successful models tend to be the simplest. Still, one wonders if a computer might somehow be programmed to determine the functional form of interatomic forces with minimal human input.

This tantalizing possibility is currently being explored using genetic algorithms [69]. By randomly generating and exchanging functional elements (mathematical operations like addition and multiplication) and selecting the fittest individuals (assessed by predictions of important energies), a population of potentials evolves until an superior form emerges. Unfortunately, it seems that left to its own devices the computer cannot even find a reasonable pair potential. It turns out that "directed" genetic algorithms, which fill in portions of a human-engineered functional template, are required for the method to be successful.

Although the directed genetic algorithm approach may ultimately be fruitful, the focus must clearly be on the theoretical direction. The same may be said of the guess-and-fit approach reviewed earlier. In the next two chapters, we shall develop two methods for obtaining information about interatomic forces directly from experimental or *ab initio* data: analysis of elastic constants and inversion of cohesive energy curves. These powerful analytic techniques have roots in the literature of solid state physics preceding the recent flurry of activity in designing interatomic potentials. With some essential innovations, we shall see that these methods can unequivocally select amongst competing intuitive ideas and provide much needed constraints on the functional form of interactions in covalent solids.

Chapter 3

Elastic Constant Relations

In 1914 I had published a paper on diamond. I assumed two kinds of forces, a radial force between two nearest neighbours and an angular force involving three neighbours. Therefore, I had two independent atomic constants. But the (cubic) crystal has three elastic constants; therefore the theory provided one relation, namely $4C_{11}(C_{11} - C_{44})/(C_{11} + C_{12})^2 = 1$. At the time no measurements of the elastic constants of diamond existed. I had to wait 31 years. Then, in 1945 in Edinburgh, I learned about new supersonic methods to measure elastic constants. Remembering the old formula, I wrote to my friend Franz Simon... and suggested to him to put one of his pupils on to this problem. Before I had finished this letter, the postman brought me my mail which included a paper by the Indian physicist Bhagavantam. It contained just these three measurements. Inserting his figures into the formula I obtained instead of 1 on the right hand side, the value 1.1 – quite a satisfactory confirmation.... This paper started off a series of investigations...establishing relations between macroscopic constants by making simple, natural approximations about the lattice forces.

– Max Born [70]

A useful theoretical approach to guide the development of potentials is to predict elastic properties implied by generic functional forms and compare with experimental or *ab initio* data. Recently, this method has been pursued by only a few authors [71, 72, 21], but apparently they did not search for elastic constant elastic constant relations implied by simple functional forms, which is our approach. This tool for understanding interatomic forces dates back to the 19th century, when St. Venant showed that the assumption of central pairwise forces supported by Cauchy and Poisson implies a reduction in the number of independent elastic constants from 21 to 15 [73]. The corresponding six dependencies, given by the single equation $C_{12} = C_{44}$ if atoms are at centers of cubic symmetry, are commonly called the Cauchy relations [73, 74]. They provide a simple test for selecting which materials can be described by a pair potential [11, 75]. Once it was realized that the Cauchy relations are not satisfied by the experimental data for semiconductors, a number of authors in this century, led by Born [9, 70], derived generalized Cauchy relations for noncentral forces in the diamond structure [10, 11].

Born's ingenious idea was to consider an *underdetermined* model (with fewer degrees of freedom than the number of independent elastic constants) and derive the implied elastic constant relations. If these relations are not satisfied, then the *functional form* cannot reproduce the data, no matter how it is fit. If they are satisfied, then we have compelling evidence that the functional form is correct. This kind of information is rare in the field of interatomic potentials; usually the validity of a functional form can only be assessed by fitting experience, which is time-consuming and inconclusive.

In this chapter we analyze the elastic properties of several general classes of many-body potentials in the diamond and graphitic crystal structures in order to gain insight into the mechanical behavior of sp^3 and sp^2 hybrid covalent bonds, respectively. These high symmetry atomic configurations must be accurately described by any realistic model of interatomic forces in a tetravalent solid. We only discuss results for three-body cluster potentials, ignoring bond order cluster functionals in the interest of simplicity. In this work the so-called Valence Force Field (VFF) models [50, 52], which can only describe small distortions of the diamond lattice, are also not considered. The goal in the VFF approach is to reproduce lattice dynamics as accurately as possible, paying little or no attention to broader transferability (with some exceptions [51, 52]). Over forty such potentials have been produced for Si alone [52], with the most recent displaying superb agreement with experiment for elastic constants and phonon frequencies [53]. However, as described earlier, unifying themes of this thesis are simplicity and transferability, and hence our motivation is quite different from VFF.

3.1 Taylor Expansion of the Cohesive Energy

The generic form of a three-body cluster potential is,

$$E = \sum_{i,j} \phi(R_{ij}) + \sum_{i,j} \sum_{k>j} \psi(R_{ij}, R_{ik}) h(l_{ijk}), \tag{3.1}$$

where we adopt the notation of Chapter 2. We make no assumptions about the functions ϕ, ψ and h (aside from differentiability, of course). Without loss of generality, we assume symmetry of the three-body radial function under exchange of atoms, $\psi(r_i, r_j) = \psi(r_j, r_i)$, which implies that $\psi_1 = \psi_2$ and $\psi_{11} = \psi_{22}$, where subscripts 1 and 2 on ψ indicate partial derivatives with respect to the first and second arguments. A natural class of symmetric radial functions is separable (like SW), $\psi(r_1, r_2) = g(r_1)g(r_2)$, but we do not require separability in this analysis. A trivial extension of the present model is to add more three-body terms of the same form, but with different angular and radial functions, as in the potential of Biswas and Haman [41]. This simply involves summation over all the three-body terms in all elastic constant formulae, so we shall not mention it again.

Now let us consider infinitesimal strains of a reference crystal structure. Assuming lattice symmetry and strain homogeneity, the energy per atom of the crystal is equal to the energy of a central atom, justifying the notation, $r_i = R_{0i}$ and $l_{ij} = l_{0ij}$. Using elementary calculus, the change in energy per atom for a three-body potential, expanded to second order in strain, is,

$$\begin{aligned} \Delta E &= \sum_i \left(\phi' \Delta r_i + \frac{1}{2} \phi'' \Delta r_i^2 \right) \\ &\quad + \sum_i \sum_{j>i} \left[h \left(\psi_1 (\Delta r_i + \Delta r_j) + \frac{1}{2} \psi_{11} (\Delta r_i^2 + \Delta r_j^2) + \psi_{12} \Delta r_i \Delta r_j \right) \right. \\ &\quad \left. + h' \Delta l_{ij} \left(\psi + \psi_1 (\Delta r_i + \Delta r_j) \right) + \frac{1}{2} h'' \psi \Delta l_{ij}^2 \right], \end{aligned} \tag{3.2}$$

where primes indicate derivatives, summation is over first neighbors, and all functions are evaluated in the unstrained, equilibrium state. The dependence of the quantities Δr_i and Δl_{ij} on strain must be computed separately for each crystal structure, which is tedious, but straight-forward. Therefore, we leave these unenlightening details to Appendix A, where the deformed bond lengths and angles are calculated for the independent strains for the diamond and graphitic lattices. Once these formulae are substituted into Eq. (3.2), the coefficients of linear terms must vanish, yielding equilibrium conditions, and those of the quadratic terms are the elastic constants.

The general form we are considering here has eight degrees of freedom for elastic properties: ϕ', ϕ'', $h\psi_1$, $h\psi_{11}$, $h\psi_{12}$, $h'\psi$, $h'\psi_1$ and $h''\psi$. In the cases of the SW and KP potentials, this number is reduced to two, ϕ'' and $h''\psi$, since $\phi' = h = h' = 0$. The Tersoff potentials, have many more degrees of freedom for elastic properties under the usual circumstances, and hence we will not be able to derive any implied dependencies. Elastic constant formulae for the related bond order potentials [21] and angularly-dependent embedded-atom potentials [72] have been calculated. These environment-dependent models have enough degrees of freedom to fit all the elastic constants (although none has managed to do so for Si, while preserving other important properties). Since our aim is to derive elastic constant relations, however, we will only consider simpler, underdetermined models.

3.2 Diamond sp^3 Hybrid Covalent Bonds

Due to the cubic symmetry of the diamond structure, there are only three independent elastic constants, C_{11}, C_{12} and C_{44}. The bond lengths and angles in Eq. (3.2) are related to the independent strains ε_1, ε_2 and γ_4 in Appendix A. After applying the equilibrium conditions, elastic constant formulae are derived by comparing with the definition,

$$V_d \Delta E = \frac{1}{2} C_{11}(\varepsilon_1^2 + \varepsilon_2^2) + C_{12}\varepsilon_1\varepsilon_2 + \frac{1}{2} C_{44}\gamma_4^2, \tag{3.3}$$

where $V_d = a^3/8$ is the volume per atom, which converts to the standard units of pressure. (a is the lattice constant.)

3.2.1 First Neighbor Interactions

With first neighbor interactions, the condition for equilibrium is

$$\phi' + 3\psi_1 h = 0, \tag{3.4}$$

which reduces the number of degrees of freedom to seven. If $h = 0$ (for the tetrahedral angle), then the pair potential must have a minimum at the first neighbor distance. At first we do not allow any internal relaxation. This has not effect on C_{11} and C_{12}, but C_{44} will be replaced by the unrelaxed value C_{44}^o. It turns out that the diamond elastic constants with a first-neighbor, three-body potential are:

$$\begin{aligned} V_d C_{11} &= K + \frac{64}{27}\psi h'', & (3.5) \\ V_d C_{12} &= K - \frac{32}{27}\psi h'', & (3.6) \\ V_d C_{44}^o &= K + \frac{32}{81}\psi h'' - \frac{16}{9} r^2 \psi_{12} h - \frac{32}{27} r \psi_1 h + \frac{16}{27}\psi h', & (3.7) \end{aligned}$$

where K is the bulk modulus,

$$V_d K = \frac{4}{9} r^2 [\phi'' + 3h(\psi_{11} + \psi_{12})], \tag{3.8}$$

and $r = \sqrt{3}a/4$ is the first neighbor distance. As a check, K can be derived from uniform dilation, and the relation $K = (C_{11} + 2C_{12})/3$ is satisfied. In the case of the

SW potential ($h = h' = 0$), we recover the formulae of Cowley, who also calculated phonon frequencies [71].

Requiring crystal stability places restrictions on the possible values of the degrees of freedom. Corresponding to the three independent modes of deformation [75], there are three inequalities,

$$\phi'' + 3h(\psi_{11} + \psi_{12}) > 0, \tag{3.9}$$

$$2\psi h'' - 9r^2\psi_{12}h - 6r\psi_1 h + 3\psi h' > 0, \tag{3.10}$$

$$\psi h'' > 0, \tag{3.11}$$

which stabilize the diamond lattice against uniform dilation ($K > 0$), simple shear of a cubic face ($C_{44}^o > 0$) and the second shear mode ($C_{11} - C_{12} > 0$), respectively. Internal relaxation can only lower the shear modulus ($C_{44} < C_{44}^o$), so the inequality of Eq. (3.10) can be strengthened using the formula for C_{44} derived below. In the important case, $h = 0$ and $\psi > 0$, the stability relations reduce to requirements of positive curvature for the pair potential and angular function, $\phi'' > 0$ and $h'' > 0$.

The only way to make our three-body cluster potential underdetermined for elastic properties is to assume $h = h' = 0$, like the SW and KP potentials. This general case is equivalent to a simple model of diamond elasticity proposed by Harrison in his Ph.D. thesis 30 years before SW [11, 12]. The Harrison model has two separate degrees of freedom, C_0 and C_1, for radial and angular forces [76], respectively, defined by the valence force field equation of Musgrave and Pope [77],

$$\Delta E = \frac{1}{2}\sum_i \frac{1}{2} C_0 \frac{\Delta r_i^2}{r^2} + \sum_i \sum_{j>i} \frac{1}{2} C_1 \Delta\theta_{ij}^2, \tag{3.12}$$

where the leading factor of 1/2 avoids double counting bonds and the sums are over first neighbors only. In terms of our cluster potential formalism, the Harrison force constants are,

$$C_0 = 2r^2\phi''(r), \tag{3.13}$$

$$C_1 \quad = \quad \frac{8}{9}\psi(r,r)h''(-1/3), \tag{3.14}$$

which follows by comparison of Eqs. (3.2) and (3.12), using $\Delta l^2 = (8/9)\Delta\theta^2$. One way to determine C_0 and C_1 is to reproduce the experimental values of C_{11} and C_{12}. In that case, the Harrison force constants for Si are $C_0 = 55.0$ eV and $C_1 = 3.2$ eV. The ratio C_0/C_1, 17.2 in the case of Si, is of the same order of magnitude in most tetrahedral semiconductors, indicating that radial forces are generally about ten times larger than angular forces [76].

In the early literature on elastic forces, unrelaxed elastic moduli were ignored, because they are not experimentally accessible. With the advent of *ab initio* calculations that predict elastic constants to within a few percent of experimental values, we can now analyze unrelaxed elastic properties as well. Since the Harrison model has two degrees of freedom for the three unrelaxed elastic constants, there is an implied relation,

$$4C_{11} + 5C_{12} = 9C_{44}^o, \tag{3.15}$$

which appears not to have been noted in previous studies (probably due to its experimental inaccessibility). As shown in Table 3.1, *the experimental and* ab initio *elastic constants for silicon satisfy this relation* within experimental and computational error. No other known elastic constant relation for covalent solids is satisfied with such precision, which is comparable to the nearly-perfect Cauchy relation $C_{12} = C_{44}$ in ionic solids [75]. This result unambiguously selects the Harrison model to describe diamond elasticity without internal relaxation. We also have an answer to one of our fundamental questions: it is indeed possible to perfectly reproduce a nontrivial manifold on the Born-Oppenheimer energy surface for covalent solids with a simple empirical potential.

On the other hand, more general cluster potentials and functionals, including the Tersoff, BH and PTHT formats, do not require our relation, and appear to be unable to satisfy it under the usual circumstances. This is demonstrated in Table 3.1 and explains why it has proven difficult to obtain good elastic properties with the Tersoff

	EXPT	LDA	SW	BH	T2	T3	DOD	PTHT	TB
C_{11}	1.67		1.617	2.042	1.217	1.425	1.206	2.969	1.45
C_{12}	0.65		0.816	1.517	0.858	0.754	0.722	2.697	0.845
C_{44}	0.81		0.603	0.451	0.103	0.690	0.659	0.446	0.534
C^o_{44}		1.11	1.172	1.049	0.923	1.188	3.475	2.190	1.35
α_B	1.07		1.11	1.03	0.33	0.28	0.71	0.93	0.11
α_H	1.16		1.00	0.98	2.99	2.31	1.69	1.71	2.80
α_{new}	0.99		1.00	1.67	1.10	0.89	0.27	1.29	0.82

Table 3.1: Comparison of elastic constants (in Mbar) for diamond cubic silicon computed with empirical models and the experimental (EXPT) or *ab initio* (LDA) values. The values for EXPT are from Simmons and Wang [78], for LDA from Nielsen and Martin [79] for tight-binding (TB) from Bernstein and Kaxiras [80] and for the empirical potentials Biswas-Haman (BH), Tersoff (T2, T3), Dodson (DOD) and Pearson-Takai-Halicioglu-Tiller (PTHT) from Balamane *et al* [19]. The Stillinger-Weber (SW) values are calculated with the analytic formulae of Cowley [71] and scaled to set the binding energy to 4.63 eV [19]. In the lower half of the table, the Born [9], Harrison [12] and new elastic constant relations are tested by calculating the ratios $\alpha_B \equiv 4C_{11}(C_{11} - C_{44})/(C_{11} + C_{12})^2$, $\alpha_H \equiv (7C_{11} + 2C_{12})C_{44}/3(C_{11} + 2C_{12})(C_{11} - C_{12})$ and $\alpha_{new} \equiv (4C_{11} + 5C_{12})/9C^o_{44}$.

potential. Note that the T3 angular function[38], which was specifically optimized for elastic properties, resembles the SW angular function, while the other versions do not. Moreover, only the T3 angular function agrees with the theoretical predictions of the quantum-mechanical bond order expansion [21, 67]. These results suggest that it may be important for the angular function to at least approximately have a minimum of zero at the tetrahedral angle within the Tersoff format too. We shall return to this issue in the next section, when we consider a quantum mechanical approximation related to the Tersoff potential.

Let us now turn to the third experimental elastic constant, C_{44} with internal relaxation. Since the Harrison model has only two degrees of freedom for this case too, it is interesting to see if it can reproduce more elastic properties. By lattice symmetry, the only possible relaxation of the two interpenetrating FCC lattices for a yz (γ_4) shear is to squeeze them together in the x direction. The driving force for internal relaxation is the resistance to stretching of sp^3 bonds, so following Kleinman we parameterize the relaxation by moving one of the two basis atoms by $a\zeta\gamma_4/4$ in the positive x direction. If $\zeta = 1$, then all bond lengths are unchanged at first order in the strain, and if $\zeta = -1/2$ the angles are unchanged, as shown in the Appendix A [81, 76]. The shear constant of the Harrison model is

$$V_d C_{44}(\zeta) = \frac{4}{9}(1-\zeta)^2 r^2 \phi'' + \frac{32}{81}(1+2\zeta)^2 \psi h'', \tag{3.16}$$

where the value of ζ is determined by minimization ($dC_{44}/d\zeta = 0$),

$$\zeta = \frac{9r^2\phi'' - 16\psi h''}{9r^2\phi'' + 32\psi h''} = \frac{C_{11} + 8C_{12}}{7C_{11} + 2C_{12}}, \tag{3.17}$$

which can be viewed as another elastic constant relation implied by the Harrison model. This one, however, is not satisfied quite so well. Using the most recent experimental value $\zeta = 0.72 \pm 0.04$ [82], the ratio of the left to the right hand side of Eq. (3.17) is 0.74, indicating that the predicted ζ is about 25% too small with the Harrison model. Earlier studies report values of ζ between 0.62 and 0.75 (which are all consistent in light of the error bounds) [83, 84, 82]. With these values, the discrepancy with the Harrison model may be as small as 14%, but in any case the model clearly somewhat underpredicts ζ.

Substituting Eq. (3.17) into Eq. (3.16), we obtain a formula for C_{44} with the Harrison model. Since there are still only two degrees of freedom for the three experimental elastic constants, there is another implied relation, originally derived by Harrison [12],

$$(7C_{11} + 2C_{12})C_{44} = 3(C_{11} + 2C_{12})(C_{11} - C_{12}). \tag{3.18}$$

The experimental elastic constants satisfy the Harrison relation to within 16% (not quite as well as the Born relation). In contrast, notice once again that the Tersoff format potentials, T2, T3 and Dodson (DOD) [44], are far from satisfying this relation, as shown in Table 3.

The performance of the Harrison model is impressive, since it is quite underdetermined for elastic properties. It provides a reasonable description of five elastic properties (C_{11}, C_{12}, C_{44}, C_{44}^o and ζ) with only two degrees of freedom. Let us now see how far the functional form can be pushed. Combining Eqs. (3.18) and (3.15), we arrive at a relation involving all four elastic constants,

$$C_{44}^o - C_{44} = \frac{(C_{11} + 8C_{12})^2}{9(7C_{11} + 2C_{12})}, \tag{3.19}$$

that expresses the effect of internal relaxation. In the case of Si, if the two degrees of freedom in the Harrison model are used to reproduce the experimental values of C_{11} and C_{12}, and thus also C_{44}^o by Eq. (3.15), then the predicted value of C_{44} from Eq. (3.19) is 0.71 Mbar, which is only 12% smaller than the experimental value of 0.81 Mbar. This explains the surprising fact [19] that the SW potential gives one of the best descriptions of elastic properties in spite of not having been fit to any elastic constants. We conclude that it is the superiority of the simple SW functional form that gives the desirable properties, not a complex fitting procedure.

Using analytic expressions for the elastic constants it is possible to devise a simple prescription to achieve good elastic properties with the Harrison model [76]. As a simple consequence of $h = 0$, the curvature of the pair potential is given by,

$$\phi''(r_d) = \frac{3V_d}{4r_d^2}(C_{11} + 2C_{12}). \tag{3.20}$$

The curvature of the angular function can be related to the second shear modulus,

$$g(r_d)^2 h''(-1/3) = \frac{9V_d}{32}(C_{11} - C_{12}). \tag{3.21}$$

Using the *ab initio* data for Si in Table 3.1, the right hand sides of Eqs. (3.20) and (3.21) evaluate to 8.1 eV/Å^2 and 5.7 eV, respectively. This provides a simple two-step procedure to maintain good elastic behavior while fitting any potential reducing to the Harrison model near the diamond lattice: (i) scale the pair interaction $V_2(r)$ to obtain the correct bulk modulus $K = (C_{11} + 2C_{12})/3$, and (ii) scale the three-body energy to set the second shear modulus. As shown above, this will lead to perfect unrelaxed elastic constants and only a 12% error in C_{44} for Si. The structural relaxation will not be as accurate, with $\zeta = 0.529$, smaller than the experimental values by 15–30%, but the overall elastic properties are still excellent, much better than might be expected *a priori* from such a simple functional form.

3.2.2 Quantum-Mechanical Interpretation

With our new results, we can better understand the successes and failures of the Harrison model, at least in the case of Si. The most important result is that the Harrison model reproduces all the unrelaxed elastic constants, not just C_{11} and C_{12} as is generally believed based on its two degrees of freedom. Therefore, the discrepancy in satisfying the Harrison relation with the experimental elastic constants is due to an inadequate description of C_{44} with internal relaxation. We may conclude that the model does not respond correctly to changes in the atomic environment that occur during relaxation. Chelikowsky's dangling bond vector, defined in Eq. (2.5), is a convenient variable to control such environment dependence because it vanishes (for any strain) in the absence of internal relaxation.

These findings have important implications for the theory of covalent bonding in solids. A natural connection between our empirical models and a quantum mechanical treatment of the electrons is supplied by the Bond Order Potential (BOP) expansion of the total energy within the Tight Binding (TB) model [67, 20, 62, 63], through the elasticity analysis of Alinaghian, Nishitani and Pettifor [21]. The BOP angular dependence for σ-bonding,

$$\begin{aligned} h_\sigma(\theta) &= a + b\cos(\theta) + c\cos(2\theta), && (3.22)\\ a &= 1 - b - c, && (3.23)\\ b &= \frac{2p_\sigma}{(1+p_\sigma)^2}, && (3.24)\\ c &= \frac{p_\sigma^2}{(1+p_\sigma)^2}, && (3.25) \end{aligned}$$

is controlled by a single quantum mechanical quantity, $p_\sigma = pp\sigma/|ss\sigma|$, which is the ratio of TB hopping matrix elements between p orbitals pointing toward each other ($pp\sigma$) and between spherically-symmetric s orbitals ($ss\sigma$) centered on neighboring atoms [21]. If $p_\sigma = 0$, the angular function is completely flat, $h_\sigma(\theta) = 1$, consistent with the nondirectionality of a pure s bond. In order to better understand the BOP angular function for $p_\sigma > 0$, let us cast it in a more transparent form,

$$h_\sigma(\theta) = \left(\frac{p_\sigma}{1+p_\sigma}\right)^2 \left(\cos(\theta) + \frac{1}{p_\sigma}\right)^2. \tag{3.26}$$

At the other extreme, the angular function for $p_\sigma = \infty$, $h_\sigma(\theta) = \cos^2(\theta)$, has a narrow, symmetric minimum at 90°, indicating that an orbital on a neighboring atom interferes (overlaps) least with a pure p bond when it is perpendicular to the bond axis. The intermediate case, $p_\sigma = 3$, corresponds to an ideal sp^3 hybrid bond, and naturally the angular function has a minimum of zero at the tetrahedral angle[1]. Therefore, when $p_\sigma = 3$, BOP reduces to the Harrison model, as far as elasticity is concerned.

[1]The $p_\sigma = 3$ angular function, $h_\sigma(\theta) \propto (\cos(\theta) + 1/3)^2$, is actually identical to the SW angular function, but angular dependence enters the functional form differently in BOP and SW, so there is little similarity away from the diamond lattice.

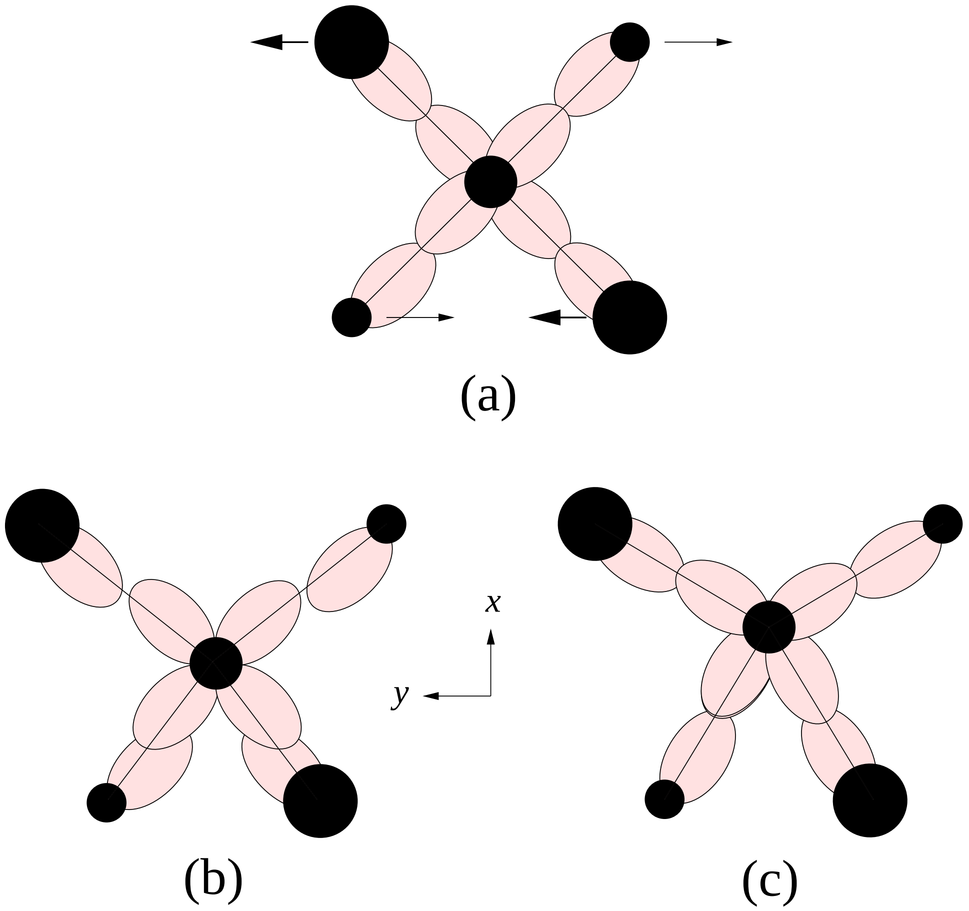

Figure 3.1: Quantum-mechanical effects during tetragonal yz (C_{44}) shear of the diamond lattice. The size of an atom (solid circle) suggests its position in the z direction out of the page. The equilibrium tetrahedron is shown in (a) with filled ellipses representing sp^3 hybrid orbitals. The strained state without relaxation is shown in (b) with rigid sp^3 hybrids, which we prove to be an accurate picture, at least for Si. The effects of internal relaxation in the x direction and rehybridization are shown in (c).

This gives us a connection with quantum mechanics: The success of the Harrison model for unrelaxed elasticity validates the Rigid Hybrid Approximation, in which electrons stay frozen in the ideal sp^3 orbitals as the lattice is deformed [76]. Harrison and Phillips showed that this approximation works well for the second shear constant $C_{11} - C_{12}$ in C, Si and Ge [86], but the implication here is that *the Rigid Hybrid Approximation accurately describes all elastic deformations without internal relaxation in Si.*

This result is somewhat surprising since rehybridization is possible even without internal relaxation [76]. For example, in the case of the C_{44} shear shown in Fig. 3.1, the content of the $|p_x>$ orbital at each atom may be reduced in two hybrids, whose mutual angle is opening, and increased in the other two, whose mutual angle is closing, with compensating changes in the $|s>$ content of each hybrid to maintain orthogonality. (The $|p_y>$ and $|p_z>$ contents will not change due to lattice symmetry.) These shifts tend to align the hybrid orbitals better with the bond axis, presumably lowering the energy of the bonding state due to the greater wavefunction overlap. At least in the case of Si, however, our results suggest that such rehybridization is minimal if relaxation of the atomic positions is suppressed.

Internal relaxation is driven by the aversion to bond stretching, which is much greater than bond bending resistance (because $C_0/C_1 \gg 1$). As shown in the figure, atomic displacements in the x direction keep bond lengths close to the equilibrium value. In the Harrison model, the relaxation distance is reached when these radial forces are balanced by angular forces that increase with the displacement. The likelihood of rehybridization is increased by relaxation, since the misalignment of the ideal hybrids is exaggerated. With the rigid hybrid interpretation given above, the inadequacy of Harrison model describing C_{44} and ζ may be seen as evidence of rehybridization during internal relaxation.

The BOP expansion may provide a convenient framework to understand this phenomenon. The transfer of probability density from $|p_x>$ to $|s>$ described above correlates with a change in p_σ. An increase in p_σ reduces the minimum of the angular function for the two bonds subtending the closing angle, and similarly, a decrease in p_σ increases the favored angle for the other two opening bonds[2]. Therefore, each bond may change its preferred angle to adapt to its environment, suggesting an *angular* driving force for relaxation that might explain why the Harrison model underestimates ζ. The angular force constant in the BOP expansion is proportional to $h''_\sigma = 2[p_\sigma/(1+p_\sigma)]^2$, so if rehybridization occurs as discussed above, the angular force constant is increased for half of the angles and decreased for the other half. The net effect of rehybridization on C_{44} is then unclear because we do not know the precise values in p_σ during relaxation, but it is certain that the rigid hybrid approximation fails.

In the elasticity analysis of Alinaghian *et. al.* a single value of p_σ without environment-dependence is used [21]. It turns out that the choice $p_\sigma = 2$ improves the prediction of C_{44} for a wide range of covalent solids compared with $p_\sigma = 3$ (see Fig. 3 of Ref. [21]). This choice lies in between the theoretically determined p_σ values of 1.57 [87] and 2.31 [76] for Si. Our results, however, imply that a BOP with $p_\sigma = 3$ should be able to fit C_{11}, C_{12} and C^o_{44} for Si. We may conclude that the quantum-mechanical parameter p_σ, and hence the angular function, should depend on the bonding environment, perhaps measured by the Chelikowsky vector, to mimic the effect of rehybridization by varying the angular force constant and the preferred angle for each bond during internal relaxation.

3.2.3 Second Neighbor Interactions

In the spirit of simplicity, let us consider extending the pair potential, but the not the three-body interaction, to include second neighbors. This already introduces two more

[2] Note that shifting the minimum of the angular function also incurs a competing penalty for the four angles per atom that do not change at leading order.

degrees of freedom, $R\phi'(R)$ and $R^2\phi''(R)$, where R is the second neighbor distance. With second neighbor forces, the pair potential does not generally have a minimum at the first neighbor distance. The new equilibrium condition is,

$$r\phi'(r) + 3R\phi'(R) + 3r\psi_1(r,r)h(-1/3) = 0. \tag{3.27}$$

Using the geometrical strain information from Appendix A, the contributions to the (unrelaxed) elastic constants from second neighbor pair interactions are ($\Delta C = C - C_{first}$):

$$\begin{aligned}
V_d\Delta C_{11} &= \frac{8}{9}r\phi'(r) + 2R\phi'(R) + 2R^2\phi''(R), && (3.28)\\
V_d\Delta C_{12} &= -\frac{4}{9}r\phi'(r) - R\phi'(R) + R^2\phi''(R), && (3.29)\\
V_d\Delta C_{44}^o &= \frac{2}{9}r\phi'(r) + R\phi'(R) + R^2\phi''(R), && (3.30)\\
V_d\Delta K &= \frac{4}{3}R^2\phi''(R). && (3.31)
\end{aligned}$$

It might seem that by symmetry we do not need to consider internal relaxation because the second neighbors are part of the same Bravais lattice (FCC) as the central atom and first neighbor relaxation has already been computed. However, this is not so for a subtle reason: due to the equilibrium condition, $\phi'(r) \neq 0$, and thus our old formulae for ζ and C_{44} are incorrect in this case.

To proceed with internal relaxation, we assume the Harrison format for first neighbor interactions with the additional radial second neighbor forces, which we shall call the H2 model. The change in shear constant of the H2 model (compared to the first neighbor Harrison model) is,

$$V_d\Delta C_{44} = \frac{2}{9}r\phi'(r)[3 + 6\zeta^2 - 2(1-\zeta)^2], \tag{3.32}$$

where the Kleinman internal strain parameter changes nonlinearly,

$$\zeta = \frac{9r^2\phi''(r) - 16\psi(r,r)h'' - 9r\phi'(r)}{9r^2\phi''(r) + 32\psi(r,r)h'' + 18r\phi'(r)}. \tag{3.33}$$

Since the H2 format has four degrees of freedom, it may be able to reproduce the four elastic constants, and the Kleinman parameter could then provide a fair test of the functional form. Since the relation of Eq. (3.15) already agrees with experiment, the deviations introduced by second neighbor forces must also satisfy it, $4\Delta C_{11} + 5\Delta C_{12} = 9\Delta C_{44}^{o}$. Solving for the values of the four parameters that exactly reproduce the four elastic constants is a nonlinear inverse problem (due to C_{44}). Harrison has apparently found the solution numerically [12], which would prove that with second neighbor radial forces the Harrison format can be extended to fit all the elastic constants.

Let us check if a simpler model with fewer parameters can still fit the data. The simplest second neighbor model we can make assumes $\phi'(R) = 0$, which we shall call the H2' model. Since we still assume $h = 0$ from the Harrison model, the equilibrium condition implies $\phi'(r) = 0$ as well. Thus, in the H2' model the pair potential is flat at both the first and second neighbor distances. The H2' model has three degrees of freedom, C_0 and C_1 from the Harrison model and a new parameter defined by,

$$V_d C_2 = R^2 \phi''(R). \tag{3.34}$$

Therefore, there is one implied elastic constant relation (which generalizes the Harrison relation), namely

$$(7C_{11} + 2C_{12} - 16C_2)(C_{44} - C_2) = 3(C_{11} + 2C_{12} - 4C_2)(C_{11} - C_{12} - C_2). \tag{3.35}$$

The third degree of freedom can be expressed in terms of the discrepancy in satisfying Eq. (3.15),

$$C_2 = \frac{1}{4}(9C_{44}^{o} - 4C_{11} - 5C_{12}), \tag{3.36}$$

which vanishes for the (first neighbor) Harrison model. Using the experimental and *ab initio* data for Si from Table 3.1, $C_2 = -0.015$, which is very small due to the nearly perfect unrelaxed elastic constants of the Harrison model. The implication is that *second neighbor forces are very weak in the diamond structure for Si.* In the particular case of the H2' model, second neighbor radial forces are 100 times weaker than angular forces, $C_2/C_1 = 0.014$, and over 1000 times weaker than first neighbor radial forces, $C_2/C_0 = 0.00031$. It is likely that other models would also predict weak second neighbor forces. This is consistent with the fact that the *ab initio* (111) stable stacking fault energy in Si, 0.005 eV/Å^2 = 0.036 eV/atom, is over 100 times smaller than the binding energy, $E_{ssf}/E_b = 0.0078$. (The atomic arrangement of first neighbors in the (111) stable stacking fault is identical to the perfect crystal).

Although second neighbor forces are small in the diamond structure, they are nevertheless important for certain materials processes. For example, the equilibrium separation of partial dislocations is determined by a balance between long-range elastic repulsion and stable stacking fault energy. So, let us continue to investigate relevance of the three constant H2' model. In addition to the elastic constant relation above, there is also a relation involving the Kleinman parameter,

$$\zeta = \frac{C_{11} + 8C_{12} - 10C_2}{7C_{11} + 2C_{12} - 16C_2}. \tag{3.37}$$

With the *ab initio* and experimental data for Si, the ratios of the left to the right hand sides of Eq. (3.35) and Eq. (3.37) are 1.16 and 0.74, respectively. In a fair test, the H2' model does not perform any better than the simpler Harrison model, thus invalidating its assumptions.

This leads us to an interesting general conclusion: adding degrees of freedom does not necessarily help fit *ab initio* data if the *functional form* is inappropriate, which is consistent with our observations about the Tersoff potentials. This also supports our suggestion in Chapter 2 that simple, theoretically motivated functional forms are superior to flexible fitting strategies. We also see the power of elastic constant relations: we are able to discard a functional form without ever having to do any fitting. Of course, this form may still be fortuitously successful for other properties, so fitting is still needed to check overall transferability. The primary value of theoretical results like elastic constant relations is in providing much needed guidance for fitting.

3.3 Graphitic sp^2 Hybrid Covalent Bonds

We can also obtain useful information about interatomic forces due to sp^2 hybrid bonds from the elastic moduli of the graphitic structure. In the following analysis, we neglect interplanar interactions, which are insignificant compared to the covalent bonds within a single, hexagonal plane. For example, the out-of-plane elastic constants C_{13} and C_{33} in graphite (carbon) are two orders of magnitude smaller than the in-plane constant C_{11} [88]. An isolated, hexagonal plane has two independent elastic constants, C_{11} and C_{12} with units of energy per unit area (and $C_{66} = (C_{11} - C_{12})/2$) [85]. Since a hexagonal plane is only a hypothetical bonding state for Si, too large in energy to be observed experimentally, we must perform *ab initio* calculations to obtain the elastic constants.

3.3.1 *Ab Initio* Elastic Constants for Graphitic Silicon

For reasons given above, it is sufficient for our purposes to consider an isolated hexagonal plane. Our *ab initio* calculations involve density functional theory in the local density approximation (LDA) using a plane wave basis with a 12 Ry cutoff and 1296 points in the full Brillouin zone for reciprocal space integrations. (These choices guarantee sufficient accuracy.) In order to preserve periodic boundary conditions, the out-of-plane lattice parameter is fixed at $c = 5.5$ Å, which is large enough to ensure neglible interplanar forces. As shown in Fig. 3.2, great care must be exercised in locating the parabolic regime of energy versus strain, to which the elastic constants (curvatures) are very sensitive. In each case, the region of linear elasticity is identified (typically strains less than 3%) using the χ^2 statistic, to measure goodness of parabolic fit. The best fit parabola is then used to approximate the elastic constant. Our results are $C_{11} = 1.79$ Mbar and $C_{12} = 0.51$ Mbar.

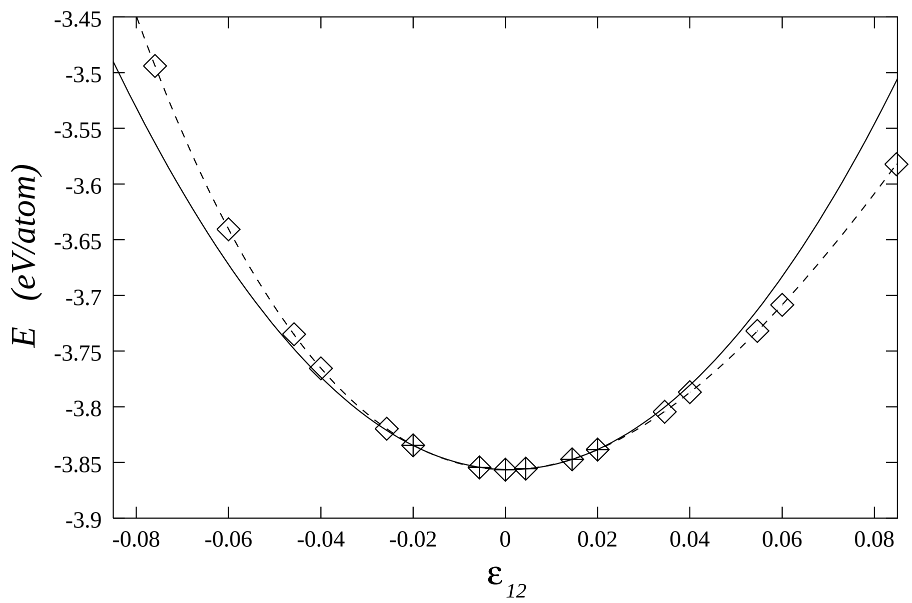

Figure 3.2: *Ab initio* (LDA) data points (diamonds) for energy versus strain of stacked hexagonal planes with $c = 5.5$ Å. The dashed line is a fit of all the data with a sixth order polynomial. The solid line is a parabola fit to the linear elastic points (+).

3.3.2 First Neighbor Interactions

Using the geometric formulae from Appendix A, we can compute the elastic constants of three-body potentials as we did for the diamond case. Since there are only two nonzero elastic constants, we should keep the functional form as simple as possible, and thus we restrict ourselves to only first neighbor interactions. The equilibrium condition for the hexagonal plane is,

$$\phi' + 2h\psi_1 = 0, \tag{3.38}$$

and using it to cancel terms, the elastic constants are:

$$\begin{aligned} A_h C_{11} &= \frac{9}{8}(r^2\phi'' + 2r^2\psi_{11}h + r^2\psi_{12}h - r\psi_1 h') + \frac{9}{16}\psi h' + \frac{27}{32}\psi h'', && (3.39)\\ A_h C_{12} &= \frac{3}{8}r^2\phi'' + \frac{3}{4}r^2\psi_{11}h + \frac{15}{8}r^2\psi_{12}h + \frac{9}{8}r\psi_1 h' - \frac{15}{16}\psi h' - \frac{27}{32}\psi h'', && (3.40)\end{aligned}$$

where $A_h = a_h^2\sqrt{3}/4$, a_h is the in-plane lattice constant and all quantities are evaluated at the equilibrium bond length $r_h = a_h/\sqrt{3}$ and angle $l = -1/2$. Consistent with symmetry requirements, $C_{44} = \frac{3}{2}r\phi' + 3r\psi_1 h = 0$, for any fist neighbor, three-body potential. However, the vanishing value of C_{44} only occurs at equilibrium, and there are still nonzero radial and angular distortions for the γ_4 shear (all at second order).

As in the diamond case, let us consider the additional simplification, $h = h' = 0$, only now these conditions apply to the *hexagonal angle* $120°$. We are effectively applying the Harrison model to three-fold coordinated sp^2 bonds. This idea, proposed by Kaxiras and also suggested by Khor and Das Sarma [45], is discussed in greater detail in Chapter 5, and for graphitic elasticity we call it the K model. The K model is justified by the BOP expansion in the Rigid Hybrid Approximation, since sp^2 hybrids have $p_\sigma = 2$ and the BOP angular function, $h_\sigma(\theta) = (4/9)(\cos(\theta) + 1/2)^2$, has a minimum of zero at the hexagonal angle in that case.

Like the Harrison model for diamond, the K model has two degrees of freedom, $r^2\phi''$ and $\psi h''$. Since there are only two elastic constants, we do not get a dependency relation, but by requiring crystal stability (equivalently, $h'' > 0$), we obtain the inequality, $C_{11} > 3C_{12}$. Using our *ab initio* data for Si, $C_{11} - 3C_{12} = 0.26$ Mbar > 0. Thus, the K model survives a test of its validity. Consistent with this result, the success of the Harrison format for sp^3 bonds makes it seem reasonable to use the same approach of the K model for sp^2 bonds, providing a unified view of covalent bond bending and stretching in the most common hybridizations.

A much more stringent test of the K hypothesis would involve calculating the elastic constant for another physically important deformation to provide a third piece of data, thus making the model underdetermined and implying a relation. For example, the hexagonal plane could be distorted out of the plane inhomogeneously, by moving one basis atom in the z direction while keeping the other fixed, which would introduce ripples into the plane resembling the ideal (111) surface. This could also be done with relaxation of the area of the plane, thus defining a generalization of the Kleinman parameter that would quantify the balance between bond-bending and bond-stretching forces in a graphitic plane. Such calculations would test the ability of the K model to describe out-of-plane bending of sp^2 bonds (to complement C_{11} and C_{12} which test in-plane bending), and would provide a means to unambiguously confirm or reject the K hypothesis for a given material. From our experience with the diamond lattice, it is likely that environment-dependence will be needed.

3.4 Comparison of sp^2 and sp^3 Bonds

The unified treatment of sp^2 and sp^3 hybrid covalent bonds with the Harrison and K models invites us to make comparisons of radial and angular stiffnesses. The relative radial stiffness is given by a simple ratio of elastic constants,

$$\frac{\phi_h''(r_h)}{\phi_d''(r_d)} = \frac{8r_d^2}{9r_h^2}\frac{A_h(C_{11}+C_{12})_h}{V_d(C_{11}+2C_{12})_d}, \tag{3.41}$$

where the subscript h refers to the equilibrium hexagonal plane with area per atom $A_h = a_h^2\sqrt{3}/4$, and d refers to the diamond lattice. For most covalent solids, the prefactor, $8r_d^2/9r_h^2$, is close to 1.0 (using the *ab initio* result for Si, $r_h = 2.23$Å, it is 0.99), so the elastic constant ratio on the right hand side of Eq. (3.41) provides a direct comparison of sp^2 and sp^3 radial forces. Our *ab initio* value of that ratio is 1.4 ± 0.1, implying that sp^2 bonds have 40% greater radial stiffness than sp^3 bonds in Si. The same result also follows directly from inverted pair potentials for the graphitic and diamond structures as described in Chapter 4.

A similar elastic analysis yields an expression for the relative angular stiffness of sp^2 and sp^3 hybrid bonds,

$$\frac{h_h''(-1/2)}{h_d''(-1/3)} = \frac{256 g_d(r_d)^2}{243 g_h(r_h)^2}\frac{A_h(C_{11}-3C_{12})_h}{V_d(C_{11}-C_{12})_d}, \tag{3.42}$$

Using our *ab initio* data for Si, we have, $g_h(r_h)^2 h_h''(-1/2)/g_d(r_d)^2 h_d''(-1/3) = 0.46 \pm 0.15$. Assuming $g_d(r) \approx g_h(r)$ with each function decreasing in accordance with inversion results (Chapter 4), then the prefactor, $256g_d(r_d)^2/243g_h(r_h)^2$, is nearly unity. In that case the ratio of elastic constants on the right hand side of Eq. (3.42) allows us to quantify the relative bending strength of the hybrid bonds. The *ab initio* value for the ratio of 0.44 ± 0.15 indicates that the angular stiffness of sp^2 bonds is smaller than that of sp^3 bonds in Si by about a factor of two, in spite of the greater radial stiffness of sp^2 bonds. This is consistent with the flatter angular function for $p_\sigma = 2$ than $p_\sigma = 3$ in the BOP expansion discussed above [21]. Our conclusion for the relative bending strength of sp^2 and sp^3 hybrids would be reversed only if $g_g(r_g)$ were smaller than $g_d(r_d)$ by at least a factor of two, which seems unlikely in light of the bond orders. No author has proposed this theoretical idea, so elastic constant analysis is leading us in a nonintuitive direction.

3.5 Conclusion

In summary, we have followed in the footsteps of Born in deriving elastic relations implied by simple functional forms of interatomic forces in covalent solids and have applied them to the case of silicon. Using *ab initio* calculations, the set of elastic constant data has been extended to include C_{44}^o for diamond and C_{11}, C_{12} and C_{44} for graphitic, which allows us to go beyond early studies restricted to experimental data (only C_{11}, C_{12} and C_{44} for diamond). The enlarged data set leads to new elastic constant relations. A surprising and important result is that the relation, $4C_{11} + 5C_{12} = 9C_{44}^o$, implied by the Harrison model is almost perfectly satisfied by the data for silicon, which is the first time such agreement has been discovered for a covalent material. We have interpreted the success of the Harrison model in terms of the Rigid Hybrid Approximation, which is apparently valid in Si for *any* elastic deformation without internal relaxation. We have also argued that weak environment-dependence, perhaps controlled by the Chelikowsky vector, is needed in the angular function to describe rehybridization. We have also suggested that second neighbor radial forces (the H2 model) can improve the Harrison model for the diamond structure, but a special case (the H2' model) having only one additional degree of freedom does not help at all.

In the case of the graphitic structure, we have found that a modified Harrison model (the K model) is consistent with the *ab initio* data for silicon, suggesting that a hybrid covalent bond (sp^3 or sp^2) is well represented by a separable, first-neighbor, three-body cluster potential whose angular function has a minimum of zero at the appropriate angle (109.47° or 120°, respectively). The strength of the angular forces is also different for the two hybrids. These results once again suggest the importance of environment-dependence in the angular function, this time depending strongly upon the *coordination*, a novel feature we shall incorporate into a model for silicon in Chapter 5. An environment-dependent angular function implies at least four-body interactions, which is consistent with quantum-mechanical predictions discussed in the previous chapter.

In addition to testing functional forms of interatomic potentials, elastic constant formulae provide quantitative guidance for fitting. Elastic constants for the diamond structure determine parameters in the potential, like the curvatures of the pair and angular functions in the Harrison model. The graphitic results allow direct comparison of sp^2 and sp^3 hybrids, which is useful in designing the environment dependence of the angular function.

Using analytic techniques and *ab initio* calculations, we have explored interatomic forces mediated by sp^2 and sp^3 covalent bonds. These results, however, say nothing of environment dependence and angular forces for more complicated structures involving overcoordination and metallic bonding (which are important for high-pressure crystal phases and the liquid). One wonders if such information could somehow be extracted directly from first principles data.

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

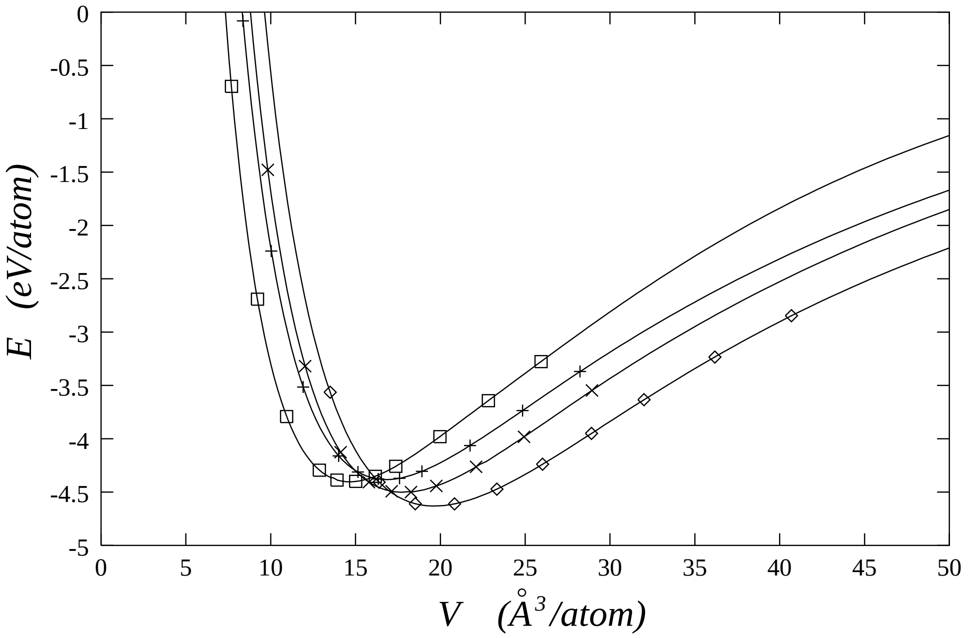

Figure 4.1: Interpolation of *ab initio* cohesive energy versus volume data for the low energy silicon crystal structures: cubic diamond (diamonds), BC8 ($\times$), BCT5 ($+$) and β-tin (squares).

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

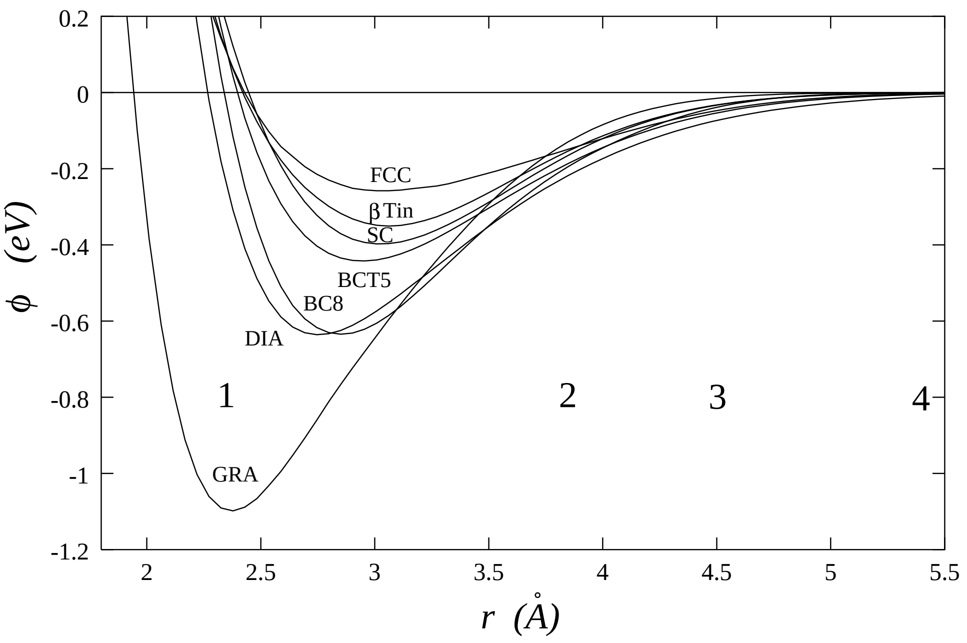

Figure 4.2: Pair potentials for silicon obtained from exact inversion of the raw *ab initio* cohesive energy versus volume curves for seven experimentally-observed (DIA, BC8, β-Tin) and hypothetical (GRA, BCT5, SC, FCC) crystal structures. Numbers indicate the positions of the first four neighbor shells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

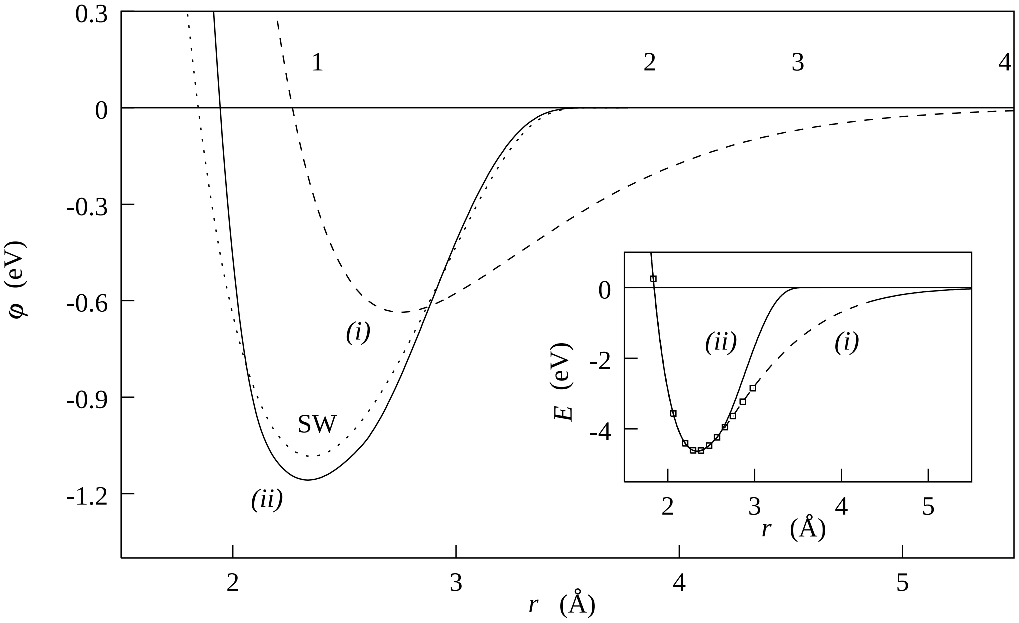

Figure 4.3: The inverted pair potential for silicon in the diamond phase (i) before and (ii) after the cutoff is imposed, compared with $\phi_{SW}(r)$ (dotted line). Numbers inside the figure indicate shell radii in the diamond lattice. The inset shows the diamond LDA data and the interpolant (i) before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

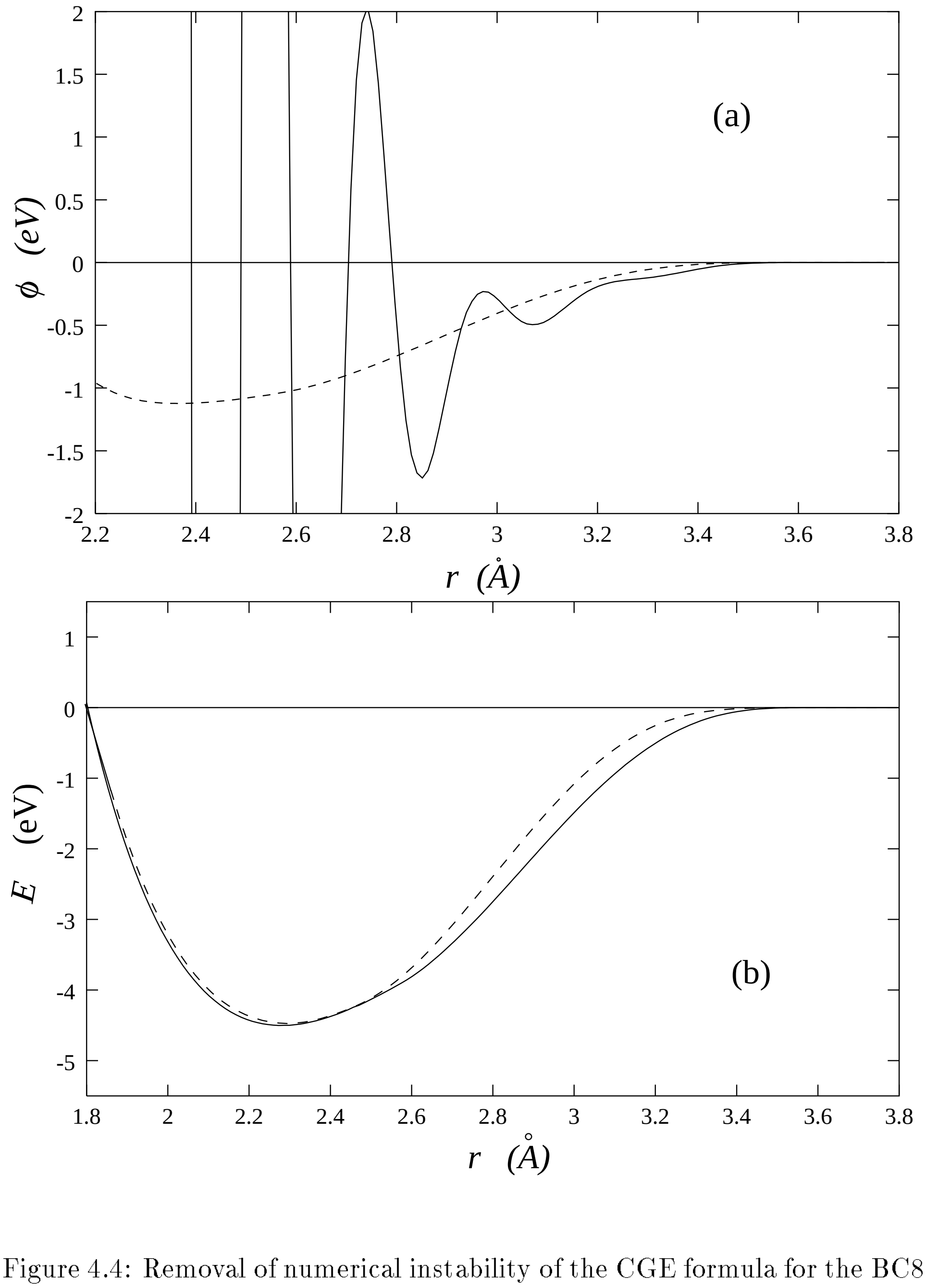

Figure 4.4: Removal of numerical instability of the CGE formula for the BC8 phase of silicon. In (a) are shown the inverted pair potential for the exact crystal structure (solid line) and the modified structure with the two nearest neighbor shells merged (dashed line). In (b), the original cohesive energy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

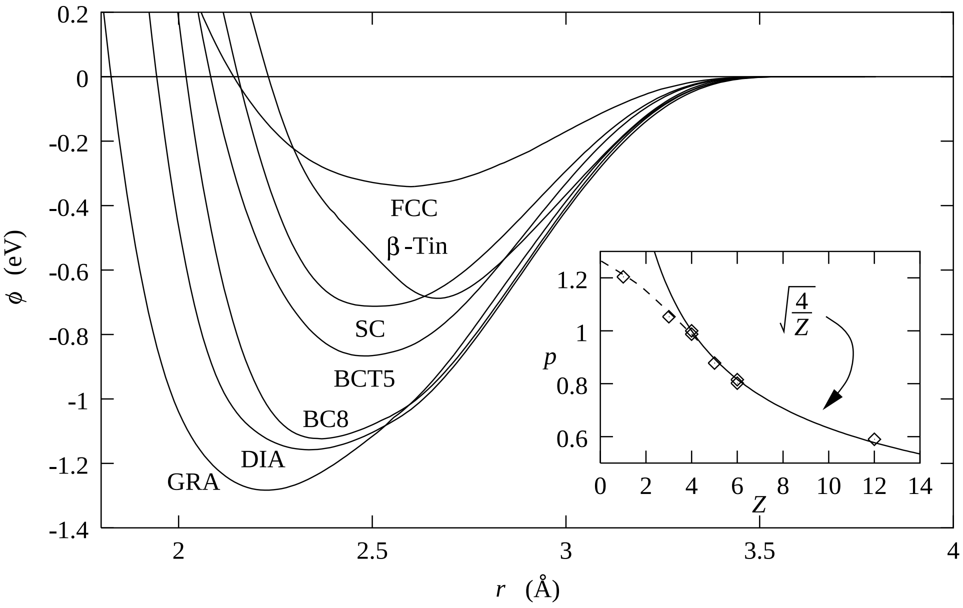

Figure 4.5: Inverted pair potentials (with cutoff) for seven silicon bulk phases. The inset shows the implied bond order p extracted from these curves (points) compared to $\sqrt{4/Z}$ (line). $p(1)$ reflects the Si_2 bond length and energy [19].

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.

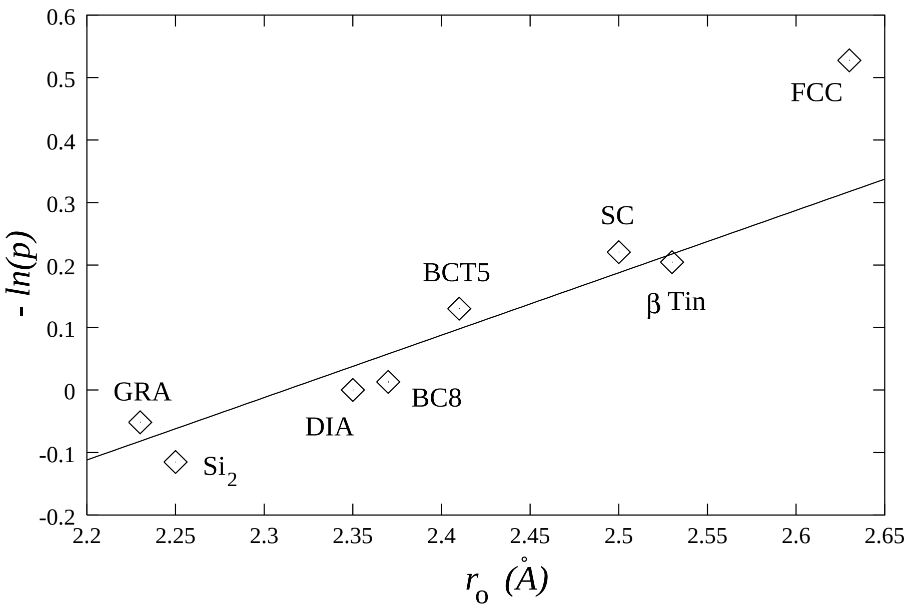

Figure 4.6: The negative logarithm of the bond order versus the radius of the minimum for our silicon inverted pair potentials. The linear fit indicates reasonable agreement with the Pauling relation between bond order and bond length.

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.

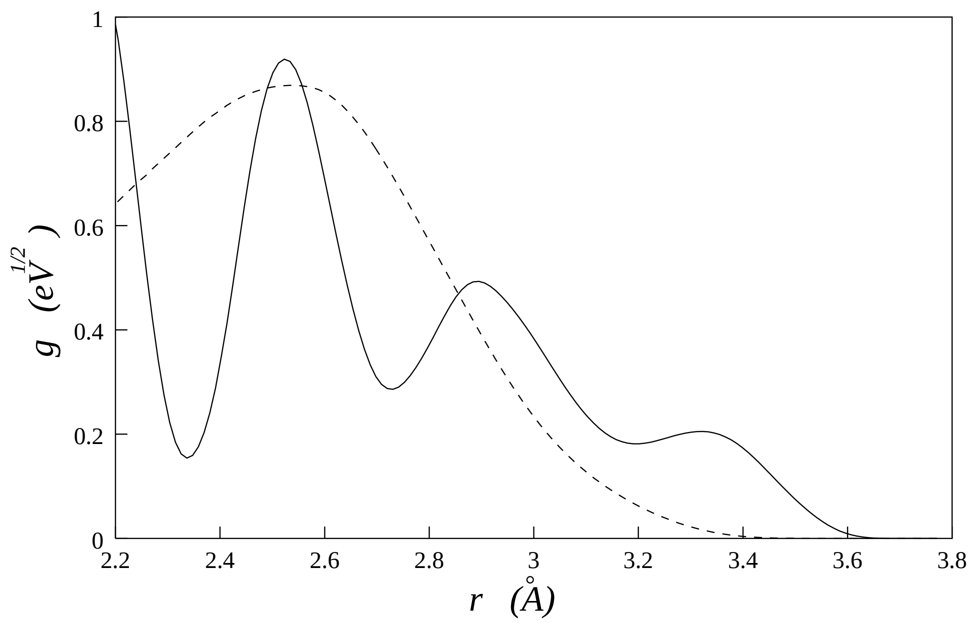

Figure 4.7: Numerical instability of three-body inversion. The inverted radial function for β-tin silicon, assuming the Stillinger-Weber angular dependence, is shown before (solid line) and after (dashed line) a cutoff function is applied to the many-body energy curve, showing how numerical instability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

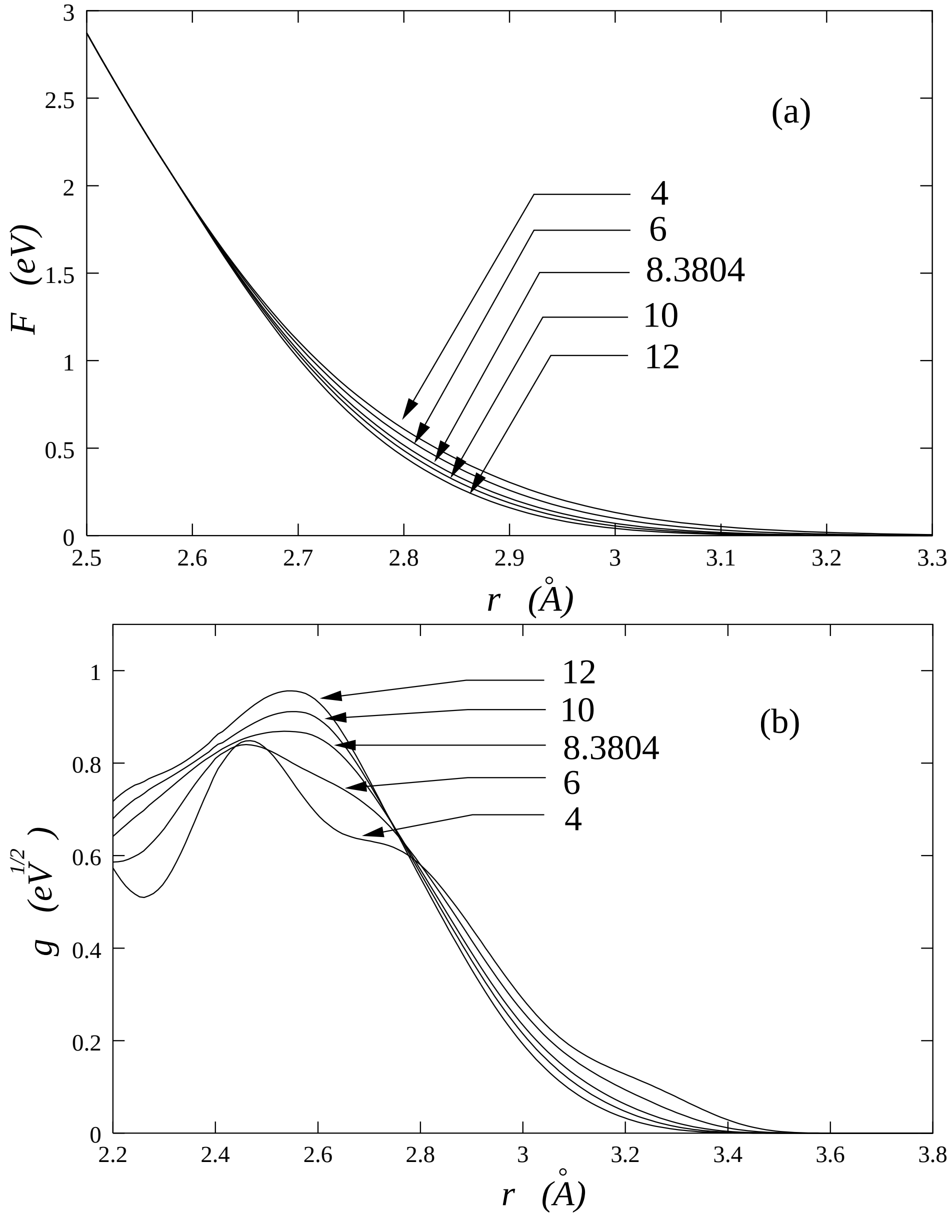

Figure 4.8: Dependence of three-body inversion on the many-body energy cutoff function. In (a), the many-body energy for β-tin silicon is shown for various values of the parameter σ_F controlling the decay of the cutoff function. In (b), the corresponding inverted radial functions are shown.

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, $\int_{a-\delta}^{a}(d^3F/dr^3)^2dr$; (*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.

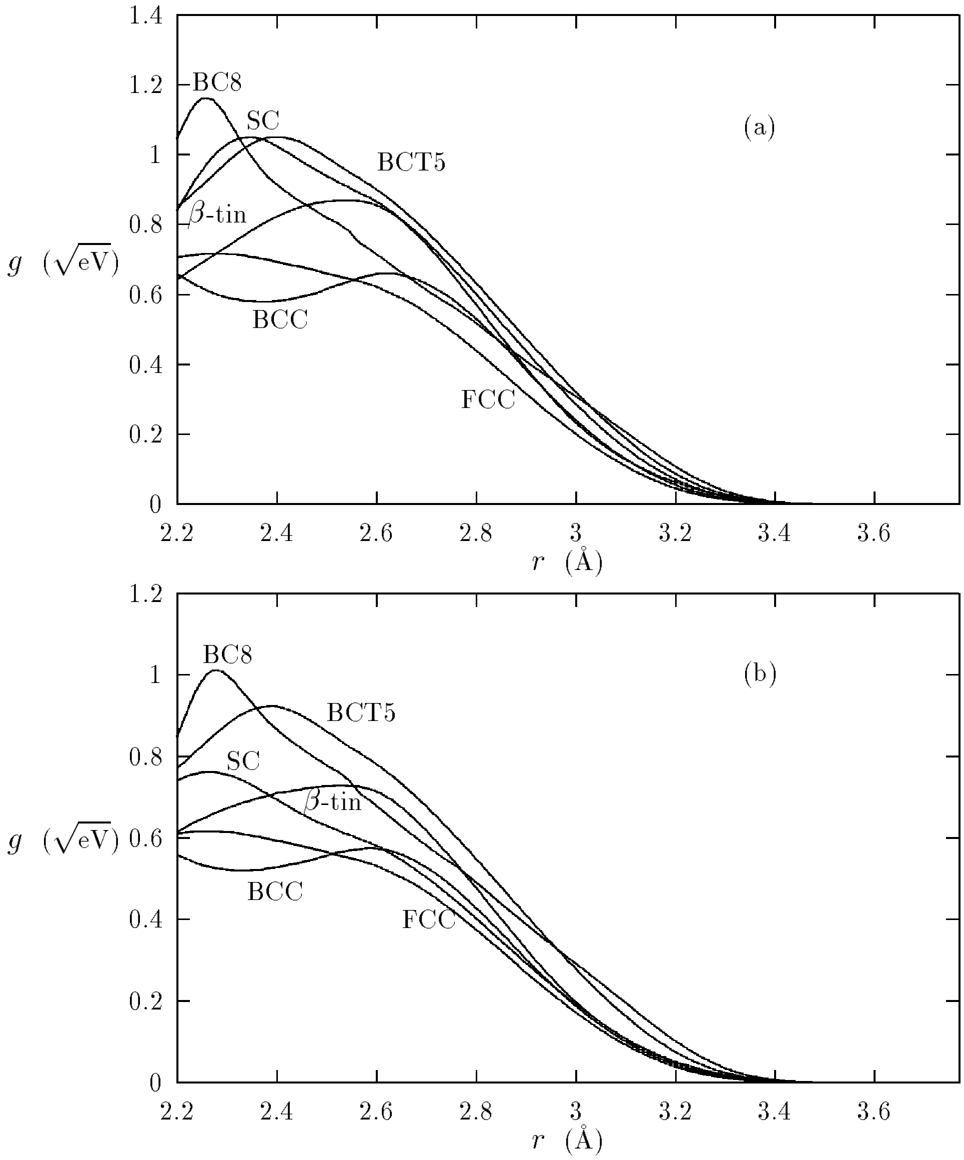

Figure 4.9: Inverted three body radial functions for silicon crystal structures, using (a) the SW angular function, $h(\theta) = (\cos\theta - \cos\theta_o)^2$, and (b) $h(\theta) = (\theta - \theta_o)^2$, where $\theta_o = \cos^{-1}(-1/3)$.

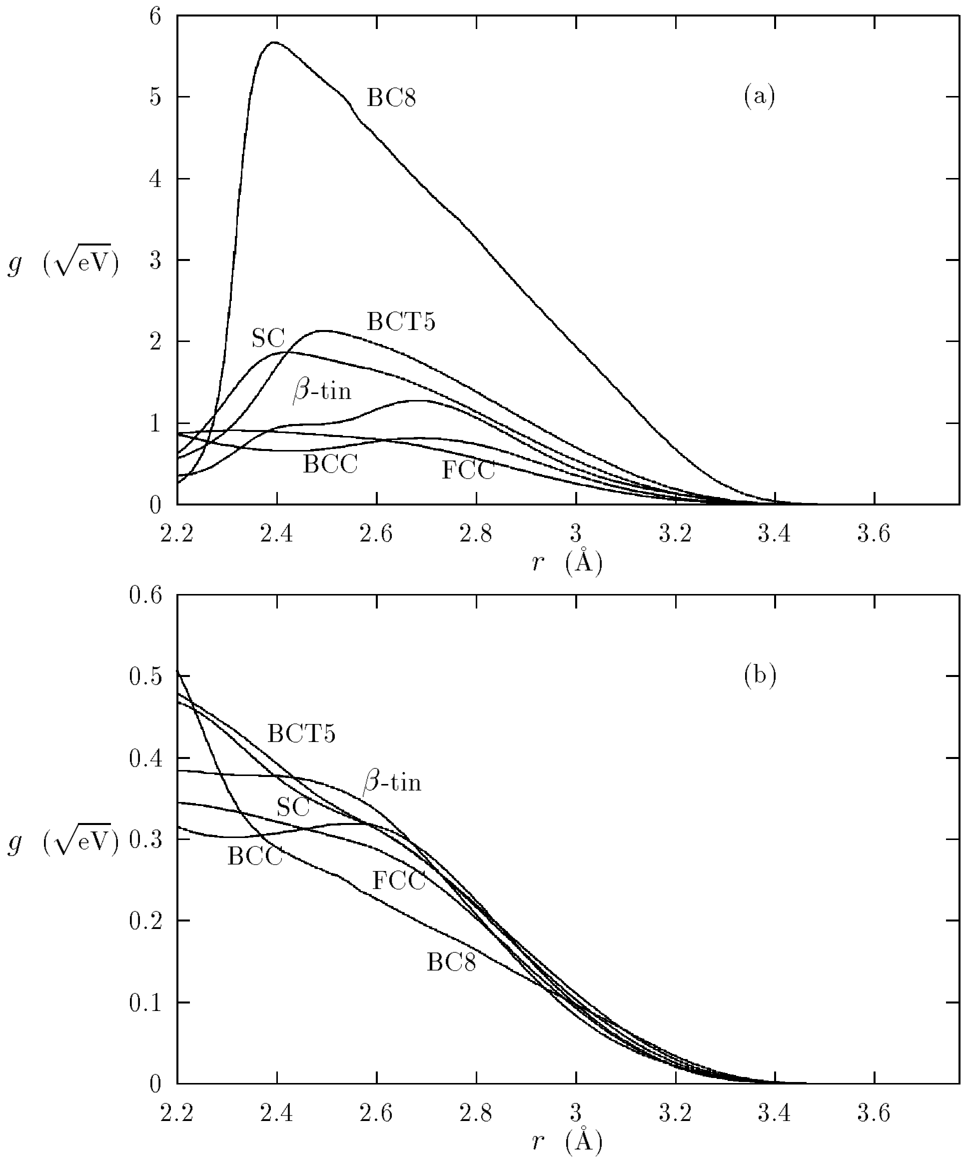

Figure 4.10: Inverted three body radial functions for silicon crystal structures, using (a) the square of the SW angular function and (b) the angular function of Kaxiras and Pandey.

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.

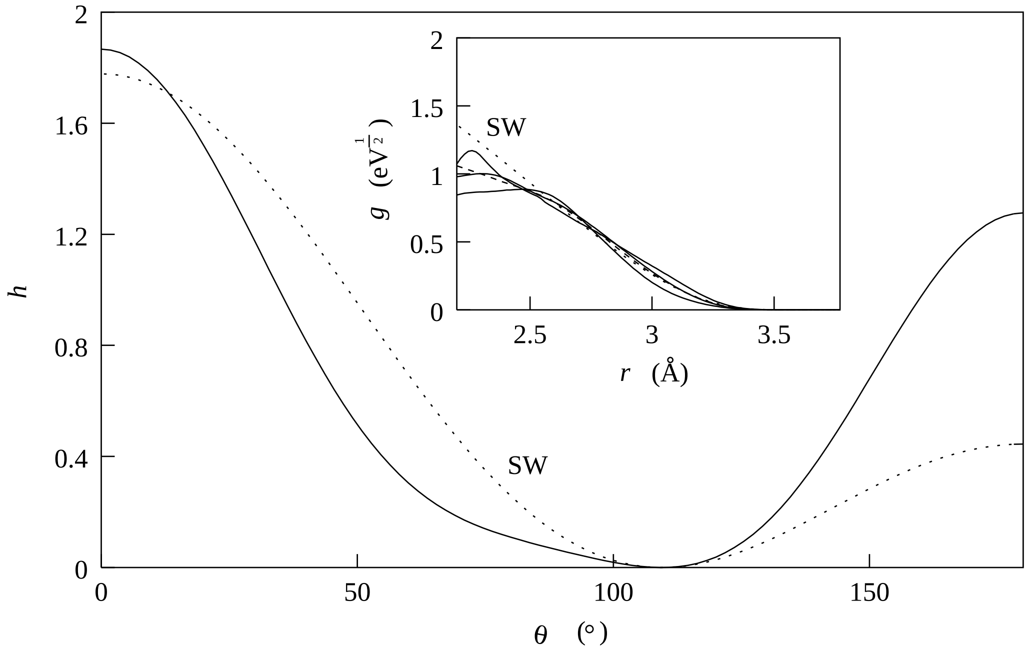

Figure 4.11: An inverted angular function for bulk silicon from the diamond, BC8, BCT5, and β-tin energy curves, compared with the Stillinger-Weber (SW) angular function. The inset shows the collapse of the inverted three-body radial functions with the average curve (dashed line) and the fitted SW radial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

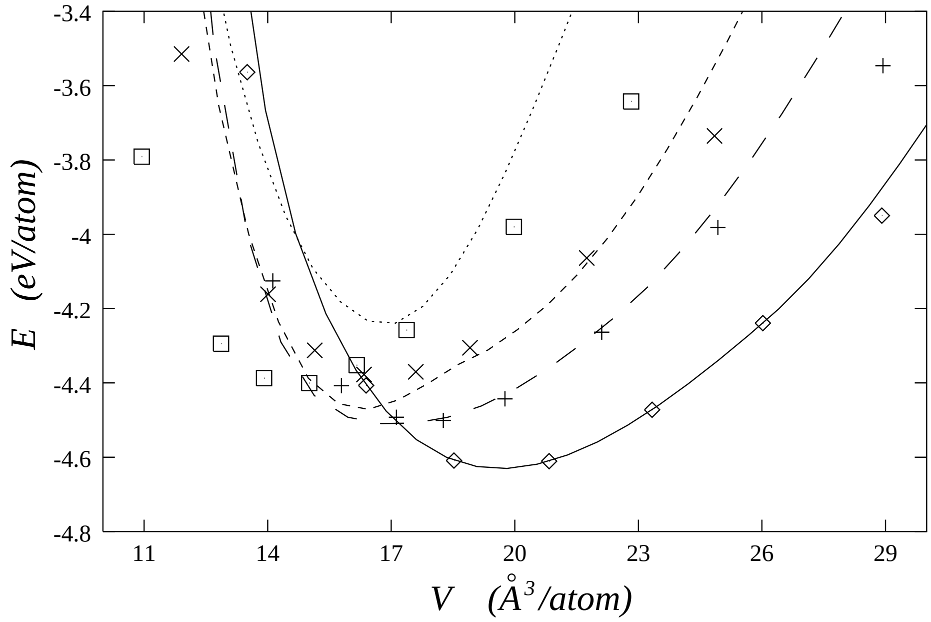

Figure 4.12: Cohesive energy curves computed with the inverted three-body potential for diamond (solid line), BC8 (long-dashed line), BCT5 (short-dashed line) and β-tin (dotted line) are compared with LDA data points for the same low energy structures, diamond (diamonds), BC8 (+), BCT5 ($\times$) and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

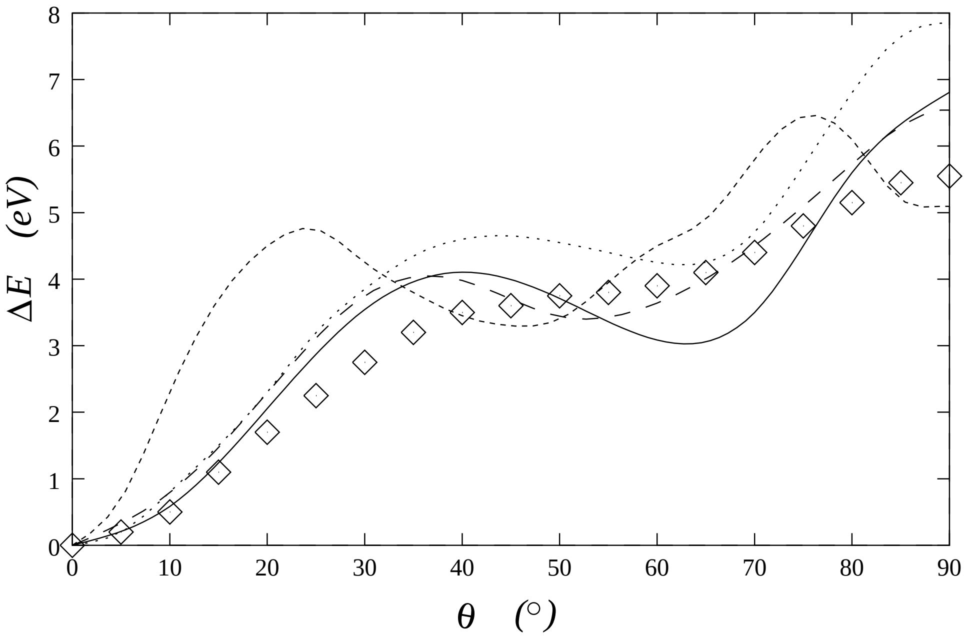

Figure 4.13: Energies of the concerted exchange mechanism for self-diffusion in silicon, as computed with the inverted potential (solid line), SW (dotted line) and T2 (short dashed line), and the EDIP potential of Chapter 5.2 (long dashed line), compared with the *ab initio* data (diamonds).

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.

Chapter 5

The Environment-Dependent Interatomic Potential

> I have restricted my work to ideal crystals though I am aware that the theory of the defects in real crystals is practically far more important. This I have left to a younger generation.
>
> – Max Born [70]

It is ironic that in answering Born's challenge to model defect structures we begin with analytic results for ideal crystals (from the preceding two chapters), some of which can be traced back to Born himself. In order to develop a model for the complex bonding in real crystals, however, a modern computer is required to search efficiently through the myriad of possible parameterizations, but not without significant human direction. Although we have seen that reasonable interatomic potentials can be derived analytically from experimental or *ab initio* data, inversion schemes become most powerful when used as theoretical guidance for fitting, for two basic reasons. The first is that inversion necessarily involves a restricted set of *ab initio* data. While the input data can be perfectly reproduced (unless it is overdetermined), it is desirable to allow an imperfect description of the inversion data in order to achieve a better overall fit of a wider *ab initio* database that includes low symmetry defect structures. The second drawback of inversion is that the class of tractable functional forms is rather limited due to issues of invertability, numerical stability, and physical validity. With the fitting approach, although there is less connection with first principles, we can explore the possibility of functional forms of greater complexity and sophistication. On the other hand, complex fitting schemes are difficult to implement; large parameter sets make it hard to judge transferability; and cumbersome functional forms obscure principles of chemical bonding and reduce the ease of force evaluation. A better approach is to incorporate the theoretically derived features of the previous chapters directly into a functional form, and then to fit the potential to a carefully chosen *ab initio* database with a minimal number of parameters. In this way, a reliable potential for bulk properties can be derived systematically, while keeping the functional form simple enough to allow for efficient computation of forces as well as intuitive interpretation of chemical bonding.

The results of this chapter are the product of almost ten years of hard work by many people, including E. Kaxiras, J. F. Justo, V. V. Bulatov, S. Yip, S. Ismail-Beigi, E. Chung and K. C. Pandey. In this chapter the current state of our empirical model for Si is presented with emphasis on the author's contributions. In Section 5.1, the theoretical results of the previous chapters are incorporated into a general functional form for interatomic forces in bulk covalent solids, called the "Environment-Dependent Interatomic Potential" (EDIP) [114], and in Section 5.2 the fitting and testing of an EDIP for bulk silicon is described.

5.1 Functional Form

5.1.1 Scalar Environment Description

Approximation of quantum models (Chapter 2.2) suggests that dependence on the local atomic environment is required to attain a transferable description of chemical bonding. In the case of silicon, these results are supported and quantitatively extended by the inversion of cohesive energy curves (Chapter 4). The simplest description of the local environment of an atom is the number of nearest neighbors. Following Tersoff [37], let us define an effective coordination number Z_i for atom i,

$$Z_i = \sum_{m \neq i} f(R_{im}) \tag{5.1}$$

where $f(R_{im})$ is a cutoff function that measures the contribution of neighbor m to the coordination of i in terms of the bond length R_{im}. The special sp^2 and sp^3 bonding geometries can be uniquely specified by their coordinations due to their high symmetry. Since environment dependence is not needed in those cases, it is natural to take the coordination number to be a constant, except when large distortions from equilibrium occur. Moreover, covalent bonds tend to involve only first neighbors, as indicated by *ab initio* charge density calculations of open structures like the diamond lattice [118]. Thus, the neighbor function is chosen to be exactly unity for typical covalent bond lengths, $r < c$, with a gentle drop to zero above a cutoff b that excludes second neighbors,

$$f(r) = \begin{cases} 1 & \text{if } r < c \\ \exp\left(\frac{\alpha}{1-x^{-3}}\right) & \text{if } c < r < b \\ 0 & \text{if } r > b \end{cases} \quad . \tag{5.2}$$

where $x = (r-c)/(b-c)$. This particular choice of cutoff function is appealing because it has two continuous derivatives at the inner cutoff c, and is perfectly smooth at the outer cutoff b. The cutoffs b and c are restricted to lie between first and second neighbors of both the hexagonal plane and diamond lattice in equilibrium, so that their coordinations are 3 and 4, respectively. The cutoff function obtained from the fitting described in the next section is shown in Fig. 5.6. Although the elastic constant analysis of Chapter 3 suggests that weak, *vector* environment-dependence is needed for even for bulk elasticity (for C_{44} only), we suppress environment dependence completely near equilibrium ($f(r) = 1$ for $r < b$) in order to achieve remarkable computational efficiency, as described in Chapter 6.

Our scalar description of the atomic environment is similar to Tersoff's, but there are notable differences. First, the perspective is that of the atom rather than the bond: With our potential, the preferences for special bond angles, bond strengths and angular forces are the same for all bonds involving a particular atom. This is in contrast to the Tersoff format [18, 37, 38, 44, 56] in which a mixed bond–atom perspective is adopted: the contribution of atom i to the strength of bond (ij) is affected by the "interference" of other bonds (ik) involving atom i. This model provides an intuitive explanation for trends in chemical reaction paths of molecules [117] and allows for both covalent and metallic bonds to be centered at the same atom, as observed, for example, in *ab initio* charge densities for the BCT5 lattice [118], which lies between the covalent diamond lattice and the metallic β-tin lattice. However, the analysis of elastic properties discussed earlier favors the present approach for environment dependence near the diamond lattice. Another important difference between our model and Tersoff's is the separation of angular dependence from the bond order. As we shall see, this allows us to control independently the preferences for bond strengths, bond angles, and angular forces in a way that the Tersoff potential cannot. By keeping the bond order simple, we can also directly use the important theoretical results that motivated the Tersoff potential in the first place.

5.1.2 Coordination-Dependent Chemical Bonding

Our potential consists of coordination-dependent two- and three-body interactions corresponding to the defining features of covalent materials: *pair bonding* and *angular forces.* The energy of a configuration $\{\vec{R}_i\}$ is a sum over single-atom energies, $E = \sum_i E_i$, each expressed as a sum of pair and three-body interactions

$$E_i = \sum_j V_2(R_{ij}, Z_i) + \sum_{jk} V_3(\vec{R}_{ij}, \vec{R}_{ik}, Z_i), \tag{5.3}$$

depending on the coordination Z_i of the central atom. The pair functional $V_2(R_{ij}, Z_i)$ represents the strength of bond (ij), while the three-body functional $V_3(\vec{R}_{ij}, \vec{R}_{ik}, Z_i)$ represents preferences for special bond angles, due to hybridization, as well as the angular forces that resist bending away from those angles. From our atomic perspective, the pair interaction is broken into a sum of contributions from each atom, and similarly the three-body interaction is broken into a sum over the three angles in each triangle of atoms. Note that due to the environment dependence, the contributions to the bond strength from each pair of atoms are not symmetric, $V_2(R_{ij}, Z_i) \neq V_2(R_{ji}, Z_j)$, if the coordinations differ.

The basic idea behind our model is that chemical bonding for an arbitrary configuration can be expressed as a simple, three-body cluster potential that adapts itself to the local atomic environment. For condensed bulk structures, the environment may be sufficiently well-described by the scalar effective coordination number. The same basic form may also be appropriate for surfaces and molecules, but a generalized, vector or matrix environment-description should be required to successfully adapt pair bonding and angular forces to highly asymmetric configurations.

Pair Bonding

As implied by the results of Chapter 4.1.9, we adopt the bond order form for the pair interaction. Drawing on the popularity of the SW potential, we use those functional forms for the attractive and repulsive interactions,

$$V_2(r, Z) = A\left[\left(\frac{B}{r}\right)^{\rho} - p(Z)\right]\exp\left(\frac{\sigma}{r-a}\right), \tag{5.4}$$

which go to zero at the cutoff $r = a$ with all derivatives continuous. This choice can reproduce the shapes of inverted pair potentials for silicon. Because we have constructed Z, and hence $p(Z)$, to be constant near the diamond lattice, our pair interaction reduces exactly to the SW form for configurations near equilibrium, thus allowing us to obtain excellent elastic properties as explained in Chapter 3. Making this choice of repulsive term with the parameters [115] obtained by the fitting to defect structures (Chapter 5.2), we can follow the procedure of Chapter 4.1 to extract the implied bond order $p(Z)$ from *ab initio* cohesive energy curves for the following crystal structures (with coordinations given in parentheses): graphitic (3), diamond (4), BC-8 (4), BCT-5 (5), β-tin (6), SC (6) and BCC (8). These structures span the full range from three- and four-fold coordinated covalent bonding in sp^2 and sp^3 arrangements, to overcoordinated atoms in metallic phases. The inverted *ab initio* bond order versus coordination is shown in Fig. 5.1, along with two additional data points. Since we have only first neighbor interactions in the diamond lattice, we can obtain another bond order for three-fold coordination from the *ab initio* formation energy (3.3 eV) for an unrelaxed vacancy. An additional data point for unit coordination comes from the experimental binding energy (3.24 eV) and bond length (2.246 Å) of the Si_2 molecule [119].

The bond order data has a clear shoulder at $Z = Z_0 = 4$ where the predicted transition from covalent to metallic bonding occurs. For overcoordinated atoms with $Z > Z_0$, the bond order approaches its rough asymptotic behavior, $p \propto Z^{-1/2}$, characteristic of metallic band structure. For coordinations $Z \leq Z_0$, the bond order departs from the $Z^{-1/2}$ divergence, due to the formation of a band gap in the LDOS associated with covalent bonds. A natural choice to capture this shape is a Gaussian,

$$p(Z) = \mathrm{e}^{-\beta Z^2}. \tag{5.5}$$

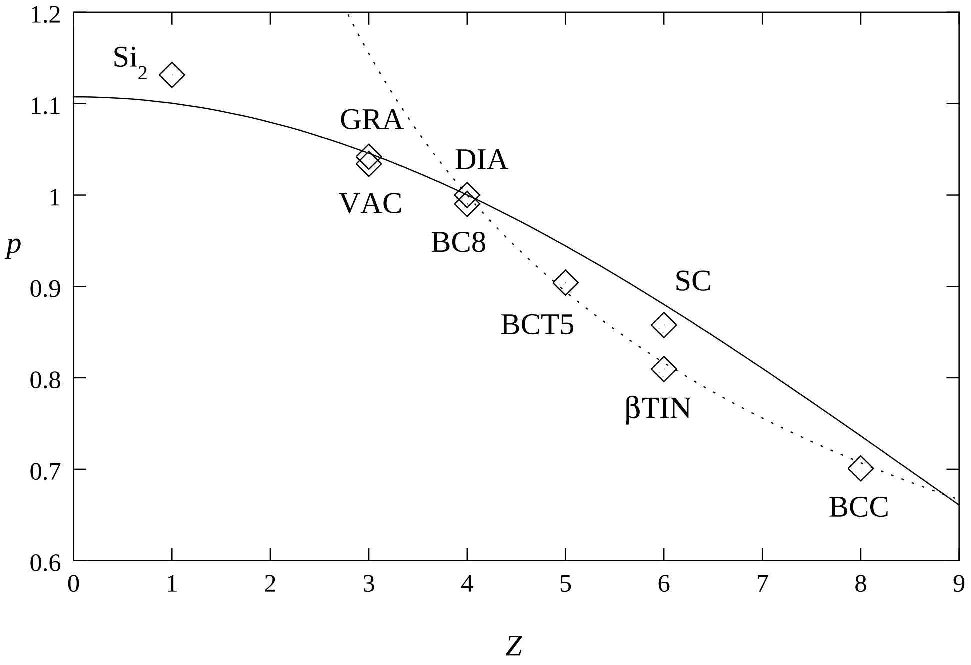

Figure 5.1: *Ab initio* values for the bond order in silicon as a function of coordination, obtained from the inversion of cohesive energy curves for the graphitic (GRA), cubic diamond (DIA), BC8, BCT5, SC, β-tin and BCC bulk structures and with additional points for the unrelaxed vacancy (VAC) and the dimer molecule (Si_2). For comparison the solid line shows the Gaussian $p(Z)$ obtained from fitting to defect structures. The dotted line shows the $1/\sqrt{Z}$ dependence, the theoretically predicted approximate behavior for $Z > 4$ (in the absence of angular forces).

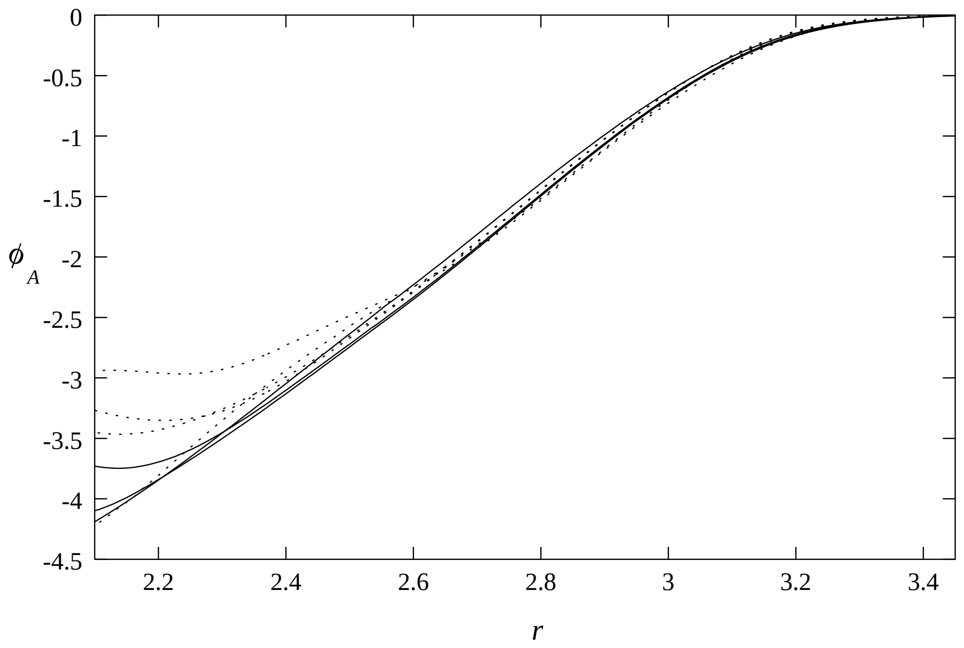

Figure 5.2: Attractive pair interactions from inversion of *ab initio* cohesive energy curves for the structures in Fig. 5.1 using the bond order and repulsive pair potential of our model. The solid lines are for the covalent structures with coordinations 3 and 4, while the dotted lines are for the overcoordinated metallic structures. The reasonable collapse of the attractive pair potentials indicates the validity of the bond order functional form of the pair interaction across a wide range of volumes and crystal structures.

In Fig. 5.1, we see that the bond order function we obtain from the fitting described in the next section is fairly close to the inversion data. It is intentionally somewhat too large for coordinations 5–8 to compensate for the small, but nonvanishing many-body energy for those structures, as explained below. The collapse of the attractive functions $\phi_A(r) = (V_2(r, Z) - V_A(r))/p(Z)$ with this choice of bond order shown in Fig. 5.2 is reasonably good, thus justifying the bond order formalism across a wide range of volumes. The deviation of the attractive functions for the metallic phases (weaker than the collapsed functions for the covalent phases) is consistent with our functional form: The pair attraction must be diminished to account for the nonvanishing (but small) many-body energy of metallic crystals. Our potential is the first to have a bond order in such close agreement with theory, which is a direct result of our novel treatment of angular forces.

Angular Forces

In a thorough comparative study of Si potentials, Balamane *et. al.* attribute the limitations of empirical models to the inadequate description of angular forces [19]. Our potential contains a number of innovations in handling angular forces, leading to a significant improvement over existing models in reproducing *ab initio* data. Analysis of elastic properties shows that, at least near equilibrium, the three-body functional should be expressed as a single, separable product of a radial function $g(r)$ for both bonds and an angular function $h(\theta, Z)$,

$$V_3(\vec{R}_{ij}, \vec{R}_{ik}, Z_i) = g(R_{ij})g(R_{ik})h(l_{ijk}, Z_i). \tag{5.6}$$

Although the radial functions could vary with coordination, in the interest of simplicity we have focused on the angular function as the most important source of coordination dependence. Inversion of *ab initio* cohesive energy curves suggests that a consistent choice for the radial functions is the monotonic SW form,

$$g(r) = \exp\left(\frac{\gamma}{r-b}\right), \tag{5.7}$$

which also goes to zero smoothly at a cutoff distance b, a value that may be smaller than the two body cutoff a. Having separate cutoffs for two and three-body interactions is reasonable because they describe fundamentally different features of bonding. Although the pair interaction might extend considerably beyond the equilibrium first neighbor distance, the angular forces should not be allowed to extend beyond first neighbors, if they are to be interpreted as representing the resistance to bending of covalent bonds.

Much of the new physics contained in our potential comes from the angular function $h(l, Z)$. Theoretical considerations suggest the following general form:

$$h(l, Z) = H\left(\frac{l+\tau(Z)}{w(Z)}\right), \tag{5.8}$$

where $H(x)$, $w(Z)$ and $\tau(Z)$ are generic functions whose essential properties we now describe. The overall shape of the angular function is given by $H(x)$, a nonnegative [17, 66] function with a quadratic minimum of zero at the origin, $H(0) = H'(0) = 0$ and $H''(0) > 0$. The theoretical studies described in Chapter 2.2 suggest a polynomial form for $H(x)$ (expansions in $l = \cos\theta$), but the exact shape is a fundamental gap in our theoretical understanding, requiring additional research. A useful tool in this regard may be direct inversion for the angular function from shear strain energy data, described in Appendix B.

Motivated by theory, the function $\tau(Z)$ is chosen to control the coordination-dependent minimum of the angular function, $l_o(Z) = \cos(\theta_o(Z)) = -\tau(Z)$, with the following form[116][1],

$$\tau(Z) = u_1 + u_2(u_3 e^{-u_4 Z} - e^{-2u_4 Z}). \tag{5.9}$$

[1] Khor and Das Sarma also used a shifted equilibrium bond angle within the Tersoff format, but they did not specify exactly how the equilibrium angle should depend upon the local environment in a general configuration [45, 46].

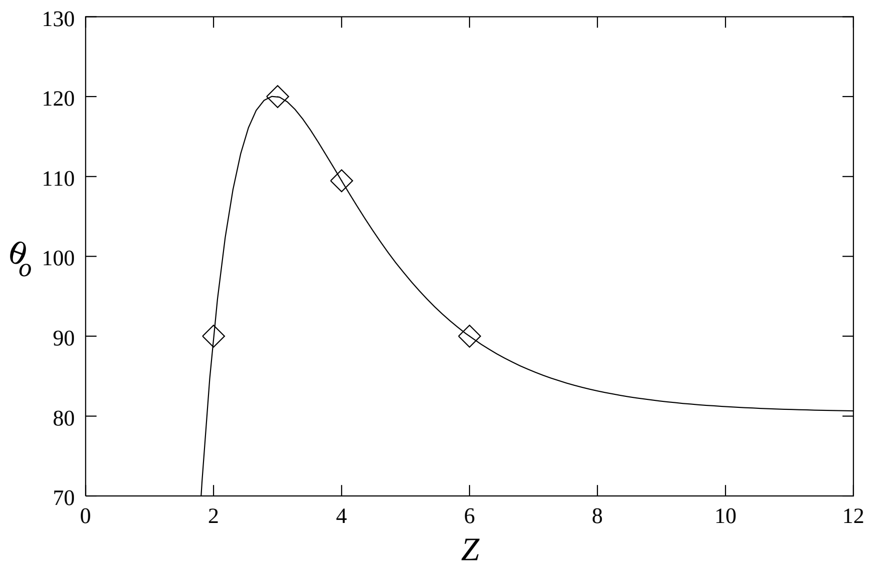

Figure 5.3: The coordination dependence of the preferred bond angle $\theta_o(Z)$ (in degrees), which interpolates the theoretically motivated points for $Z = 2, 3, 4, 6$, indicated by diamonds.

The parameters, $u_1 = -0.165799$, $u_2 = 32.557$, $u_3 = 0.286198$, and $u_4 = 0.66$, were chosen to make the preferred angle $\theta_o(Z) = \cos^{-1}[-\tau(Z)]$ interpolate smoothly between several theoretically motivated values, as shown in Fig. 5.3: We have already argued that $\tau(4) = 1/3$ and $\tau(3) = 1/2$ (so that sp^3 and sp^2 bonding correspond to the diamond and graphitic structures respectively), which determines two of the four parameters in $\tau(z)$. The remaining two parameters are selected so that $\tau(2) = \tau(6) = 0$ or $\theta_0(2) = \theta_0(6) = \pi/2$. For two-fold coordination, this choice reproduces the preference for bonding along two orthogonal p-states with the low energy, nonbonding s state fully occupied. For six-fold coordination, the choice $\theta_0(6) = \pi/2$ also reflects the p character of the bonds. However, structures with $Z = 6$ like SC and β-tin are metallic, with delocalized electrons that tend to invalidate the concept of bond-bending underlying the angular function, a crucial point we shall address shortly. The vanishing many-body energies for the graphitic plane and diamond structures allow fitting of the pair interactions $V_2(r,3)$ and $V_2(r,4)$ to be guided by Eq. (3.20), which determines $V_2''(r_d,4)$ from the bulk modulus, and Eq. (3.41), which requires $V_2''(r_h,3)/V_2''(r_d,4) \approx 1.4$. Moreover, the shifting of the minimum of the angular function in our model incorporates coordination-dependent hybridization in a way that other potentials cannot.

As an important aside, let us note that this choice of $\tau(Z)$ is not unique. It simply captures the essential feature of an environment-dependent minimum of the angular function. The BOP expansion suggests a similar (and simpler) function from Eq. (3.26): if we assume $p_\sigma = Z - 1$, then $\tau_\sigma(Z) = 1/(Z-1)$. The only major difference with the EDIP function is $\tau_\sigma(2) = 1$, indicating a favored angle of 180° for $Z = 2$, which also makes sense for *sp*-bonding. Otherwise, the EDIP function is in excellent agreement with the BOP expansion for $Z > 2$. Although we shall not pursue it here, another reasonable starting point to build a coordination-dependent, many-body potential for Si may be to let $p_\sigma = Z - 1$ in the BOP expansion.

Through the function $w(Z)$, the EDIP angular function has another novel coordination dependence to represent the covalent to metallic transition. The width of the minimum $w(Z)$ is broadened with increasing coordination, thus reducing the angular stiffness of the bonds as they become more metallic. Similarly, as coordination is decreased from 4 to 3, the width of the minimum is increased to reproduce the smaller angular stiffness of sp^2 bonds compared to that of sp^3 bonds. Thus, the function $w(Z)$ should have a minimum at $Z_0 = 4$ and diverge with increasing Z. Fitting of the model can be guided by Eq. (3.21), which determines $w(4)$ from the second shear modulus, and by Eq.(3.42), which requires $w(3)/w(4) \approx \sqrt{2}$. These features can be captured with the choice,

$$w(Z) = w_o Z^{-4\delta} \mathrm{e}^{\delta Z}, \tag{5.10}$$

where $\delta \approx 2.3$. The softening of the angular function is important because it allows the decrease in cohesive energy per atom concomitant with overcoordination to be modeled by a weakening of pair interactions. In contrast, cluster potentials like SW and the inverted potential of Chapter 4.3 penalize overcoordinated structures with increased three-body energy that overcomes the decrease in pair bonding energy. This is an unphysical feature, since overcoordinated structures do not even have covalent bonds, and the many-body energy cannot be viewed as a consequence of stretching sp^3 bonds far from the tetrahedral geometry. In this sense, the reasonably good description of liquid Si (a metal with about 6 neighbors per atom) with the SW potential appears to be fortuitous. The large overestimation of the bulk modulus of β-tin (another six-fold coordinated metal) by SW and the inverted potential is another sign of the unphysically strong angular forces.

The coordination dependence of our angular function makes it possible for the first time to reproduce the well-known behavior of the bond order. The reason is that the contribution of the three-body functional to the total energy is suppressed for ideal crystals and overcoordinated structures. The shifting of the minimum makes the three-body energy vanish identically for sp^2 and sp^3 hybrids, and the variable width greatly reduces the three-body energy in metallic structures. With the three-body energy suppressed, we can use our knowledge of the bond order for the graphitic, diamond, β-tin and other lattices from inversion of cohesive energy curves to capture the energetics of these structures in the pair interaction, as described above. Several other potentials have tried to incorporate the bond order predicted from theory, but the uncontrolled many-body energy makes it impossible to connect directly with theory. Our treatment of angular forces is intuitively appealing because the forces primarily model the bending of covalent bonds, with the control of global energetics left to the pair interactions.

5.1.3 Discussion

In summary, recent theoretical innovations have been used to arrive at a functional form that describes the dependence of chemical bonding on the local coordination number. Bond order, hybridization, metalization and angular stiffness are all described in qualitative agreement with theory. Consistent with our motivation, we have kept the form as simple as possible, reproducing the essential physics with little more complexity than existing potentials. The fitted implementation of the model described in the next section involves only 13 adjustable parameters. We have theoretical estimates of almost half of the parameters, thus greatly narrowing the region of parameter space to be explored during fitting. The remaining parameters are chosen to fit important bulk defect structures.

Considering the theory behind our model, we can anticipate its range of applicability. We have shown that the structure and energetics of the diamond lattice can be almost perfectly reproduced. Because small distortions of sp^3 hybrids are accurately modeled, we would also expect a good description of the amorphous phase. Defect structures involving sp^2 hybridization should also be well described. In general, the model should perform best whenever the coordination number can adequately specify the local atomic environment. This certainly includes sp^2 and sp^3 hybridization and some metallic states, but might also include more general situations in which atoms are more or less symmetrically distributed, like the liquid and amorphous phases and reconstructed dislocation cores and grain boundaries. The theory behind the model begins to break down for noninteger coordinations, since our effective coordination number is only a way of smoothly interpolating between well-understood local structures. More seriously, no attempt is made to handle asymmetric distributions of neighbors, which are abundant in surfaces and small clusters. Theory suggests that our model may be fitted to provide a good description of condensed phases and defects in bulk tetrahedral semiconductors, such as Si, Ge and with minor extensions perhaps alloys such as SiGe, that can be understood in terms of simple principles of covalent bonding.

5.2 Fitting and Tests for Bulk Silicon

5.2.1 Fitting to Defect Structures

After choosing the functional form, the next important selection to be made in constructing an empirical potential is the set of experimental or *ab initio* data used to determine the adjustable parameters. Just as the functional form is designed to handle particular bonding arrangements, so too must the input data be carefully chosen to include representative properties within the target range of transferability of the potential. Experience shows that it is very difficult (if not impossible) to simultaneously fit to all the important classes of bonding states. Instead, the most successful approach is to focus on a single set of structures sharing similar bonding characteristics, and then to extend the range of transferability incrementally by adding new physics to the functional form and expanding the database. In the case of silicon, we have begun by working with bulk crystal and defects, with disordered phases as the next step (still in progress). These types of structures are within the theoretically predicted range of validity of the EDIP functional form, so we may hope to obtain a successful potential from a carefully controlled fitting strategy. Our first goal is a superior potential for simulations of condensed phases and bulk defects, such as defect diffusion, plastic deformation, radiation damage, amorphous vibrations, sintering, bulk melting and solid phase epitaxial growth. Ignoring other important structures like surfaces and small clusters is reasonable because a satisfactory degree of transferability for bulk defect structures not yet been attained by any model.

The *ab initio* fitting database used to construct the latest version of EDIP for silicon [115], compiled primarily by E. Kaxiras, includes: the diamond structure (cohesive energy and lattice constant), the experimental diamond elastic constants (C_{11}, C_{12} and C_{44})[78], formation energies of unrelaxed point defects (vacancy and tetrahedral and hexagonal interstitials)[122, 123, 124], the concerted exchange (CE) path for self-diffusion (sampled at 10° intervals)[111], and selected points on the generalized stacking fault (GSF) energy surface [120, 121]. These target properties are used to optimize the 13 adjustable parameters: A, B, ρ, β, σ, a, b, c, λ, γ, α, Q_o and μ. (The last two parameters are defined below.) The fitting of the parameter vector $\{c_i\}$ to the input data $\{E_j\}$ (energies and elastic constants) is accomplished with a simulated annealing algorithm[109] that minimizes a least-squares cost function,

$$\Phi(\{c_i\}) = \sum_j \left(\frac{\tilde{E}_j - E_j}{\sigma_j} \right)^2 + P(\{c_i\}), \tag{5.11}$$

where $\tilde{E}_j$ is the value of property E_j predicted by the potential with the given parameter set, σ_j^{-1} are fitting weights (*e.g.* large for diamond cohesive energy and small for a particular CE data point), and $P(\{c_i\})$ is a nonnegative function that punishes unacceptable properties (like an energy lower than diamond or an incorrect equilibrium crystal volume). The art of fitting potentials lies in careful choice of all of these variables. Letting the annealing program run blindly always leads to frustration. Instead, one must regularly monitor the fitting routine and make adjustments to the weights, set of input properties, the penalty function, the random walk of the parameter vector, and the annealing schedule in response to the improvement or deterioration of the overall fit. Additional testing of the potential in parallel to the fitting is also required, for example, to search for spurious low-energy structures or expose trends behind poorly described structures (like large coordinations or small angles). A good fitted model is one which can survive a long barrage of attempts to invalidate it.

The parameters obtained from the latest simulated annealing fit, due to J. F. Justo in close collaboration with V. V. Bulatov, S. Yip, E. Kaxiras and the author, are given in Table 5.1. The fitted pair bonding function $V_2(r, Z)$ is plotted in Fig. 5.4 for several values of the coordination. Note the close similarity between $V_2(r, Z)$ and the inverted pair potentials for silicon shown in Fig 4.5, a built-in feature of the EDIP functional form.

The angular dependence $h(l, Z)$ is depicted in Fig. 5.5 through the three-body energy

Table 5.1: Values of the parameters that define the latest version of EDIP for bulk silicon.

A = 12.360638 eV	B = 1.6039258 Å	ρ = 1.3950202
a = 3.4557809 Å	b = 3.1640691 Å	c = 2.4504896 Å
σ = 1.3386900 Å	λ = 0.4610305 eV	γ = 0.2037403 Å
Q_o = 135.14236	μ = 0.7468472	β = 0.0063757
α = 4.0000000		

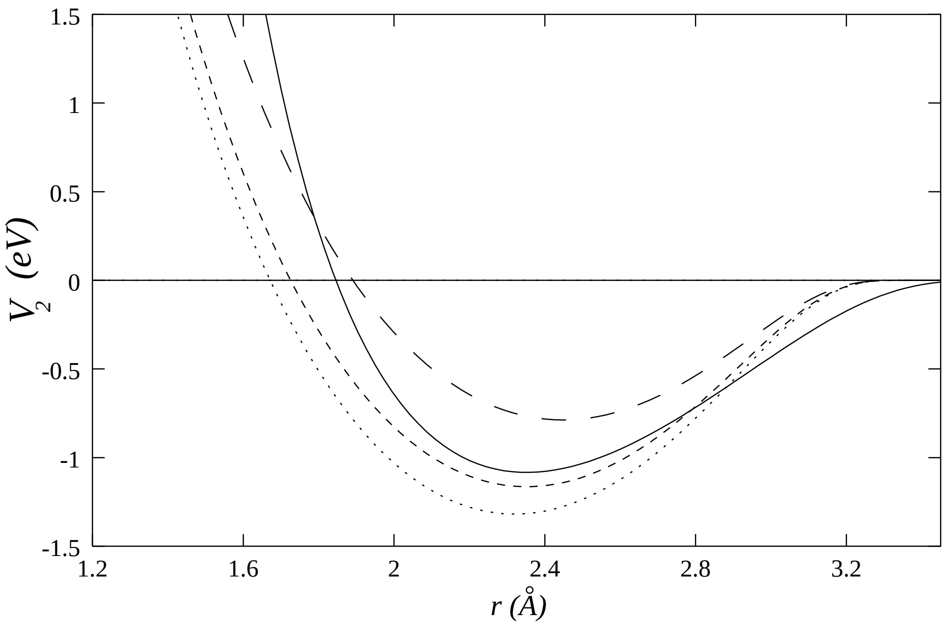

Figure 5.4: The two-body interaction $V_2(r, Z)$ as a function of separation r for different coordinations: $Z = 3$ (dotted line), $Z = 4$ (small-dash line), and $Z = 6$ (large-dash line), compared with the SW (solid line) pair interaction.

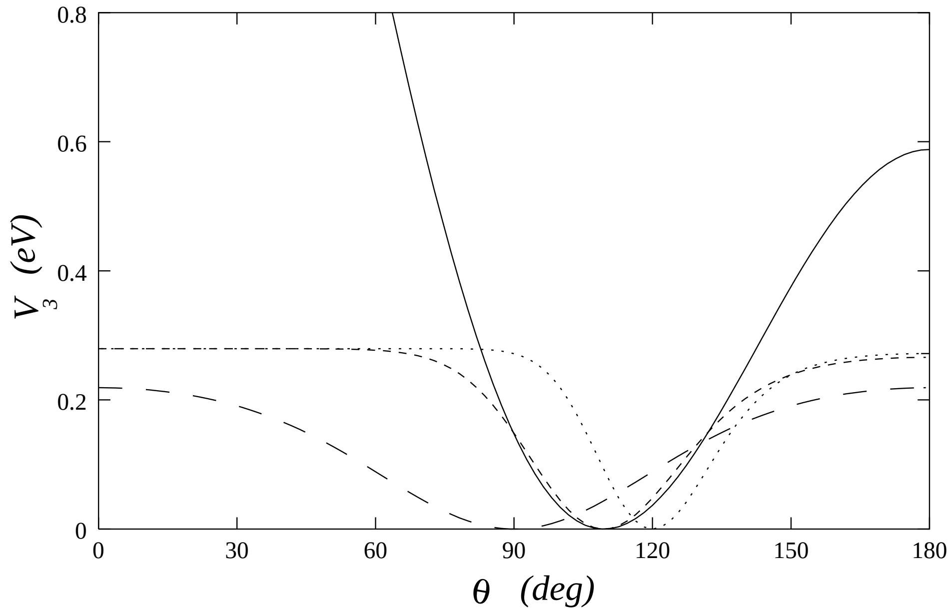

Figure 5.5: The three-body interaction $V_3(r, r, \cos\theta, Z)$ for a pair of bonds of fixed length $r = 2.35$ Å subtending an angle θ . The V_3 term is shown for several coordinations: $Z = 3$ (dotted line), $Z = 4$ (small-dash line), and $Z = 6$ (large-dash line), and compared with the SW (solid line) three-body interaction.

$V_3(r, r, l, Z)$ for a triplet of atoms with two bond lengths fixed at the equilibrium bond length $r = 2.35$ Å in differently coordinated environments. The shift of the minimum and the variable width are clearly seen, reflecting the physical trends we have already discussed. Note that for $Z = 12$ angular forces are almost completely suppressed, with $V_3 < 0.01$ eV for all angles. There is one inconsistency: the width of the angular function (greater angular forces) is smaller for $Z = 3$ than for $Z = 4$, the reverse of the theoretical prediction in Chapter 3. Unfortunately, the theoretical result came just after the last fit, but it may be incorporated into subsequent versions.

The particular choice of the shape of the angular function in the current version of EDIP for Si, due to J. F. Justo, resembles the angular dependence of the MFF potential [55]. An advantage of the MFF form over SW is its increased flexibility, containing two independent parameters instead of one. In order to make the similarity explicit, the MFF notation, $Q(Z) = w(Z)^{-1/2}$, is adopted,

$$h(l, Z) = \lambda \left[1 - \exp\left(-Q(Z)(l + \tau(Z))^2\right)\right]. \tag{5.12}$$

Although this inverted-Gaussian shape is identical to the MFF angular function, the EDIP dependence is much more sophisticated due to its environment dependence. The current choice for $Q(Z)$ is monotonically decreasing,

$$Q(Z) = Q_o \mathrm{e}^{-\mu Z}, \tag{5.13}$$

in contrast with the function postulated in Eq. (5.10), which requires that $Q(Z)$ have a maximum at $Z = 4$. Therefore, the functional form used in this version captures the covalent to metallic transition, but does not reproduce the relative bending strength of sp^2 and sp^3 hybrid covalent bonds. This may not be a serious problem, since the theoretical result comes from elastic moduli of the diamond and graphitic lattices, which is only relevant for angles near 109° and 120°, respectively. Nevertheless, the utility of the function in Eq. (5.10) is currently being explored, and, should it succeed, it might show that theory can steer the fitting process in a useful, but nonintuitive, direction.

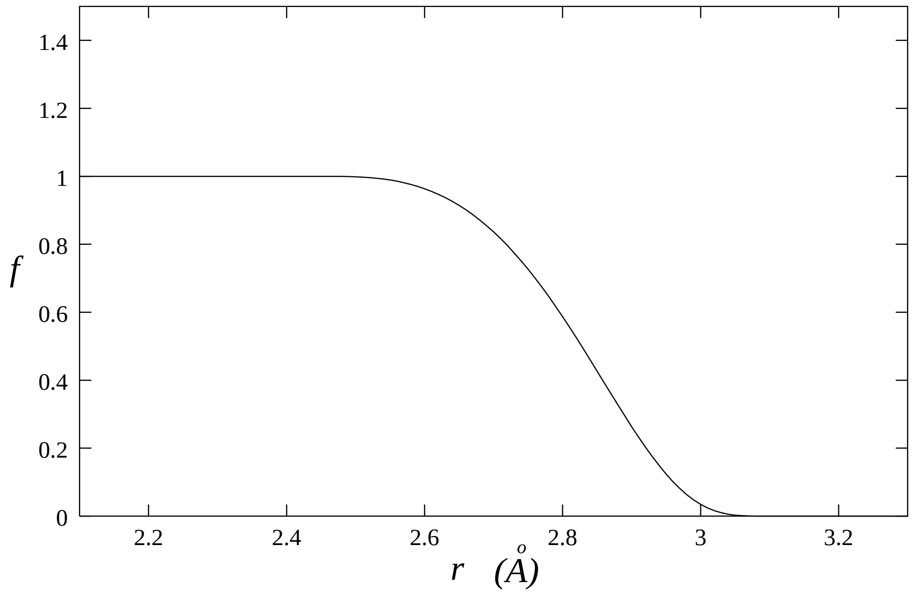

Figure 5.6: The function $f(r)$ that determines the contribution of each neighbor to the effective coordination Z.

The fitted neighbor function $f(r)$ is shown in Fig. 5.6. Note that the inner cutoff c, the largest radius for a full contribution to the coordination, is somewhat small to accurately count neighbors in closepacked structures. The coordinations measured for the *ab initio* equilibrium metallic crystals are: $Z = 5.863$ for β-tin (6), $Z = 5.994$ for SC (6), $Z = 7.801$ for BCC (8) and $Z = 11.319$ for FCC (12), where the actual coordinations are given in parentheses. Note, however, that the coordinations for all covalently bonded structures are correct, as well as for the mixed bonded BCT5. Thus, we have an accurate representation of the typical bulk atomic environments, placing the potential on firm theoretical ground.

5.2.2 Tests for Bulk Properties and Defects

Crystal Structure

The experimental binding energy, lattice constant and bulk modulus of the ground state diamond structure are perfectly reproduced by EDIP (as we enforce with large fitting

	EXPT	EDIP	SW	T3	TB
C_{11}	1.67	1.71	1.61	1.43	1.45
C_{12}	0.65	0.63	0.82	0.75	0.85
C_{44}	0.81	0.72	0.60	0.69	0.53
C_{44}^{o}	1.11	1.11	1.17	1.19	1.35
B	0.99	0.99	1.08	0.98	1.05
$C_{11} - C_{12}$	1.02	1.08	0.79	0.68	0.60
$C_{12} - C_{44}$	-0.16	-0.09	0.22	0.06	0.32

Table 5.2: Elastic constants of the diamond phase of silicon in Mbar, from experiment (EXPT) [78] (and first principles for C_{44}^{o} [79]) compared with the predictions of empirical potentials EDIP and SW (from the formulae of Cowley [71]) and T3 [19], as well as a tight binding model (TB) [80]. The dimensionless Kleinman internal strain parameter is also compared with experiment ([82, 83, 84].

weights). The cohesive energy curves of the low energy bulk phases are reasonably well described in the typical volume range for condensed phases of 15–20 Å^3, including some equilibrium volumes and energies (β-tin and BC8). There are also, however, some interesting inaccuracies in the cohesive energy curves, whose discussion we shall postpone until the next section.

Elastic Constants

The elastic constants of EDIP are given in Table 5.2 and compared with experimental and *ab initio* data, as well as the predictions of other empirical models. The overall agreement with experiment is excellent, a marked improvement over existing models, both empirical potentials and semi-empirical tight-binding Hamiltonians. This is not surprising, because, using the theoretical arguments of Chapter 3, we have built realistic elastic behavior directly into the EDIP functional form. The fitted EDIP potential reproduces the experimental values of C_{11} and C_{12} to within three percent, and therefore, as a consequence of the new elastic constant relation, Eq. (3.15), a good value of C_{44}^o is ensured. Indeed, the EDIP and *ab initio* C_{44}^o are identical. The second shear modulus $C_{11} - C_{12}$ is also very good, and the bulk modulus of EDIP is in perfect agreement with experiment. The value of C_{44} with internal relaxation is low, as it is with all empirical models, but the EDIP value is closer to experiment than the others, being only 11% too small. In contrast, with the most popular and successful version of the Tersoff potential (T2) the predicted value of C_{44} is 0.103 Mbar, almost an order of magnitude to small. The quality of elastic constants of the fitted EDIP is at the theoretical limit of its functional form, as calculated in Chapter 3. Also as predicted by that analysis, the Kleinman internal strain parameter is underestimated by EDIP, $\zeta = 0.494$, compared with the experimental, SW and T3 values, 0.74, 0.63 and 0.67, respectively. Finally, EDIP is the only transferable potential for silicon known to predict the negative Cauchy discrepancy, $C_{12} - C_{44}$, aside from the embedded-atom potential of Baskes was explicitly fit to it [48]. Even a number of semi-empirical tight-binding models cannot reproduce this important property[80], and those that can are not as accurate as EDIP [125].

In spite of its realistic elastic constants, however, EDIP does not provide a significant improvement over SW and T3 for phonon spectra[126]. The EDIP phonon frequencies are overestimated like those of the other potentials, especially along the zone boundary.

Point Defects

Point defect formation energies for vacancy (V), tetrahedral interstitial (I_T) and hexagonal interstitial (I_H), computed with EDIP are given in Table 5.3 and compared with LDA and other empirical models[2]. As an important example of an activated complex,

[2] The unrelaxed point defect energies in the fitting database, were computed using a plane wave basis with at least a 12 Ry cutoff and adequate sampling of the Brillouin zone for reciprocal space integrations.

		LDA	EDIP	SW	T3	TB
V	E_f	3.3-4.3	3.94	4.63	4.10	4.4
	ΔE_f	0.4-0.6	1.81	0.40	1.2	0.45
I_T	E_f	3.7-4.8	5.49	12.21	6.92	4.5
	ΔE_f	0.1-0.2	1.16	6.96	3.47	0.5
I_H	E_f	4.3-5.0	6.19	17.10	8.22	6.3
	ΔE_f	0.6-1.1	0.85	10.15	3.61	1.3
CE	E_f	5.5	6.55	7.90	6.50	5.5
	ΔE_f	0.9	2.51	3.26		1.8

Table 5.3: Ideal formation energies E_f^{ideal} of point defects (in eV) and relaxation energies $\Delta E = E_f^{ideal} - E_f^{relaxed}$ with EDIP using a 54 atom unit cell, compared with *ab initio* (LDA) [111, 122, 123, 124], SW and T3 [19, 42] and tight-binding (TB) [80] results.

energies for the saddle point of the concerted exchange path (CE) [111] are also given. The unrelaxed formation energies for all four defect configurations are included in the fitting database, and thus the EDIP values are fairly close to LDA. Note that the SW (and to a lesser degree T3) unrelaxed interstitial energies are much higher, indicating unphysical intolerance to overcoordinated structures with small angles, a point we have already noticed in the context of the inverted angular function of Chapter 4.3. Although the relaxed defect formation energies with SW and T3 are reasonable, they clearly fail in predicting the energy released upon relaxation from the ideal configuration. On the other hand, the EDIP relaxation energies are in fair agreement with *ab initio* calculations, in spite of only being fit to ideal structures. Note that EDIP predicts outward relaxation of the vacancy due to its first-neighbor, bond-order functional form, in analogy with the (111) 1×1 surface. Another important structure not included in the database is the split $< 111 >$ interstitial, whose (relaxed) formation energy with EDIP is 2.95 eV, compared with 4.68 eV for SW and 3.30 eV from *ab initio* calculations[127]. In agreement with *ab initio* results, EDIP predicts the split interstitial to be lowest energy interstitial configuration, while SW does not.

Finally, let us consider the concerted exchange path in more detail. Although the saddle point energies might suggest that EDIP, SW and T3 are equally good for this complicated and important path of local deformations, that is not the case. As shown in Fig. 4.13, EDIP outperforms SW, T3 and the inverted potential in describing the path overall (but keep in mind that only EDIP was fit to this data). EDIP still predicts unphysical minimum around 55°, but it is not as pronounced as with the inverted potential. SW predicts a somewhat smaller metastable minimum at 65°, but the saddle point energy is overestimated. T3 fails more seriously than any of the other models,

Supercells for the point defect calculations included 53-55 atoms, so that long range elastic relaxation energies, on the order of 0.01 eV, are ignored. For example, the formation energies for the unrelaxed defects used in the fitting are: 3.3 eV for the vacancy, 3.7 eV and 4.3 eV for the tetrahedral and hexagonal interstitials, and 5.47 eV for the concerted exchange saddle point configuration.

because there is a minimum rather than a saddle point at 90°, as well as another deep, unphysical minimum at 45°.

Extended Defects

The only extended defects in the fitting database are generalized stacking faults (GSF) of the {111} glide plane[120, 121]. In particular, three points (including energy saddles) on the glide set and three on the shuffle set are included. Generalized stacking fault energy surfaces are important as *ab initio* atomistic input to lattice-continuum models like the Peierls-Nabarro theory of dislocations, and test the ability of the potential to handle bond rupture and formation. Cross sections of the {111} glide set GSF energy surface in the high symmetry $< 11\overline{2} >$ and $< 1\overline{1}0 >$ directions computed with EDIP (by J. F. Justo [126]), SW and first principles are shown in Fig. 5.7. EDIP provides an good description of the lowest energy $< 11\overline{2} >$ cut, which passes from an ideal lattice to a stable stacking fault. The path to the stable stacking fault is more faithfully reproduced by EDIP than SW (although some points were fit). In the $< 110 >$ cut, which connects equivalent ideal crystal structures in the direction of the Burger's vector for a full dislocation, EDIP also outperforms SW. Although SW better fits the energy of the saddle point, this energy is very large and hence is physically irrelevant. In the important part of the curve ($E < 5$ eV/Å^2), EDIP is very close to the *ab initio* data points, in spite of not being fit to them.

Note that all first neighbor potentials, like SW and EDIP, predict a vanishing stable stacking fault energy. Although this may compromise some situations, such as long-range partial dislocation interactions, it is not a serious problem since the *ab initio* stable stacking fault energy is quite small, 0.006 eV/Å^2 (0.15 eV per atom on the glide plane), and the accuracy of empirical potentials is rarely better than a few tenths of an eV per atom.

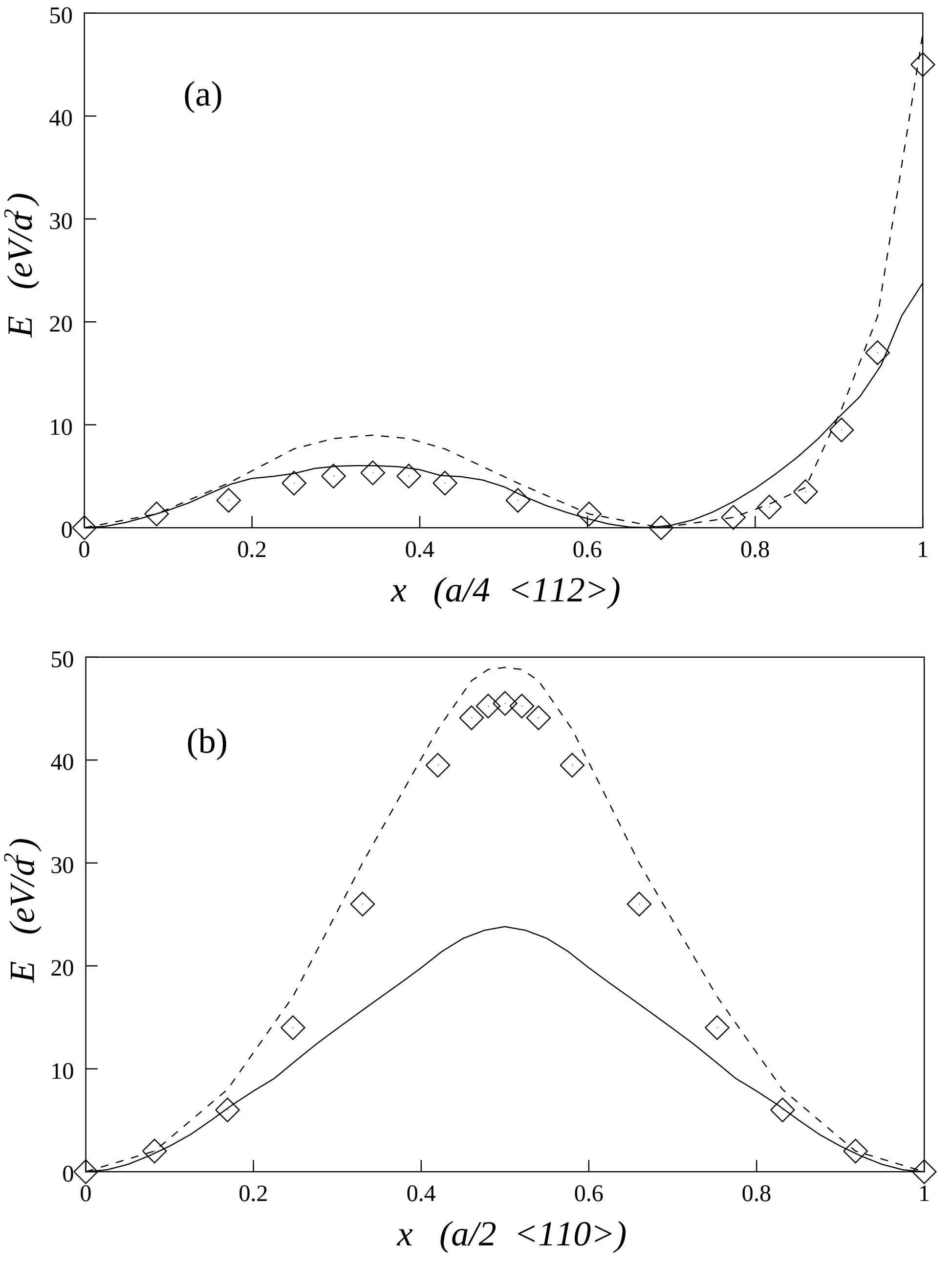

Figure 5.7: Cross section of the {111} glide set generalized stacking fault energy surface obtained from calculations using LDA (diamonds) the SW (dashed line) and EDIP (solid line) along the (a) $< 11\overline{2} >$ and (b) $< 1\overline{1}0 >$ directions.

Dislocation Cores

A much more stringent test of the transferability of EDIP comes from dislocation cores, about which no information is included in the fitting database. The lowest energy dislocations in tetravalent semiconductors like silicon are the 60° full and full screw in the $< 110 >$ / $\{111\}$ glide set slip system [128]. The 60° full lowers its energy by dissociating into 30° and 90° partial dislocations separated by a stable stacking fault, and the screw dissociates into two oppositely oriented 30° partials. It is the mobility of these partial dislocations that is believed to control plastic deformation in silicon. Due to the long-range $(1/r)$ stress fields around dislocations, any calculation of dislocation core energies or dynamics must involve a large number of atoms. Thus, partial dislocation cores in silicon provide an important materials application that is well-suited for theoretical study with empirical potentials, as long as sufficient transferability can be demonstrated to build confidence in their predictions. Unfortunately, no previous empirical potential is known to describe partial dislocations in silicon, even qualitatively.

In contrast, recent calculations with EDIP (by J. F. Justo in collaboration with V. V. Bulatov and S. Yip) have shown remarkable transferability for dislocations, making realistic simulations and energy studies possible. Here we shall only consider the highlights of core reconstructions; for a complete treatment with EDIP, including anti-phase defects, kink complexes and crossed dislocations, the reader is referred to Ref. [126]. First let us consider the 90°-partial dislocation core, which forms a simple asymmetric reconstruction by connecting dangling bonds across the core, thus lowering the energy by 0.87 eV/$\mathcal{B}$ according to first principles calculations [129], where $\mathcal{B}$ is the periodicity along the dislocation line. The SW potential cannot handle the core of the 90°-partial at all, because it does not support any reconstruction, *i.e.* the ideal, unreconstructed configuration is the stable minimum of energy [130]. The T3 potential does support reconstruction, but the predicted energy gain of 0.37 eV/$\mathcal{B}$ is too small by more than a factor of two [130]. In contrast, EDIP not only predicts the proper reconstruction, but also nearly the correct energy, 0.80 eV/$\mathcal{B}$.

Now let us turn to the 30°-partial dislocation core, which reconstructs by forming dimers to eliminate dangling bonds, much like the (100) surface. The *ab initio* reconstruction energy is 0.43 eV/$\mathcal{B}$ [131]. In this case, the SW potential predicts the correct reconstruction, but the energy 0.84 eV/$\mathcal{B}$ is too large by almost a factor of two. The T3 potential fails to describe the 30°-partial even qualitatively, because the reconstruction energy is negative[3]. Once again, in this case, EDIP not only predicts the reconstruction, but also reproduces the energy. In fact, the EDIP reconstruction energy of 0.45 eV/$\mathcal{B}$ is remarkably (and perhaps fortuitously) close to the *ab initio* value.

In summary, EDIP has attained an unprecedented degree of transferability for the silicon bulk crystal and defects. It is the first potential capable of a full description of partial dislocations with quantitative accuracy that appears to surpass even semi-empirical tight-binding models. The excellent elastic constants ensure an accurate treatment of long-range forces, and the core reconstructions and core defects are also well-described, in spite of not being explicitly fit.

Amidst these successes, however, there are hints of problems. For example, partially reconstructed metastable energy minima are predicted by EDIP along the reaction path connecting the ideal and reconstructed cores of both the 30°- and 90°-partial dislocations [126]. These minima are unphysical, but they will not affect transition rates or stability of the reconstructions during simulations. Similar oscillations in energy have also recently been observed for large shear distortions by E. Tadmor. These kinds of artificial energy changes occur when neighboring atoms pass through the cutoffs a and b, thus changing the local environment and making subtle contributions to the energy before they come close enough to form strong covalent bonds, which we have seen are quite well described by EDIP. In the setting of bulk defects, the density remains fairly con-

[3]This means that a metastable energy minimum appears at the reconstructed configuration, but it is higher in energy than the ideal structure, and hence is not favored.

stant, and typically only a small fraction of bond lengths fall near the cutoff distances during a simulation. In other cases of interest where volume changes are important, however, like melting or pressure-induced phase transitions, we may anticipate more serious problems.

5.2.3 Cohesive Energy Curves

We have paid special attention to cohesive energy curves for bulk crystal phases in building the theoretical foundation for EDIP, so it is natural to ask how the fitted potential performs. Cohesive energy curves computed with EDIP for the crystal structures we considered in Chapter 4 are shown in Fig. 5.8 (a)[4]. Unfortunately, the curves display large oscillations and bear little resemblance (across the entire volume range) to the *ab initio* curves of Fig. 4.1. Not everything is wrong: the diamond and BC8 curves are in close agreement with the *ab initio* curves across a broad range of volumes near their minima, indicating an accurate description of the lowest energy sp^3 bonding state. For volumes close to the minima of the corresponding *ab initio* curves, the EDIP curves are also not unreasonable. The obviously unphysical features lie away from those volumes, on both sides. At large volumes, every curve, even diamond, has a strange wiggle. In Fig. 5.8 (b), we see that those oscillations lie in the range where first neighbors contribute only partially to the EDIP coordination, $0 < f(r) < 1$. Another wiggle appears at smaller volumes roughly where second neighbors start to change the coordination. These volumes are not within the theoretical regime of validity for EDIP, which requires a correct determination of the coordination at the bulk density. So, these oscillations are not problematic for most bulk simulations, but there are other problems which are more serious. If the curves are recomputed with Z fixed at the value appropriate for each lattice, as shown in Fig. 5.9, then we see that the oscillations at high density are

[4]Note that these curves are constructed by dilating the *ab initio* structures. Allowing internal relaxation with EDIP might bring down some energies, but the general trends and oscillations should be insensitive to relaxation.

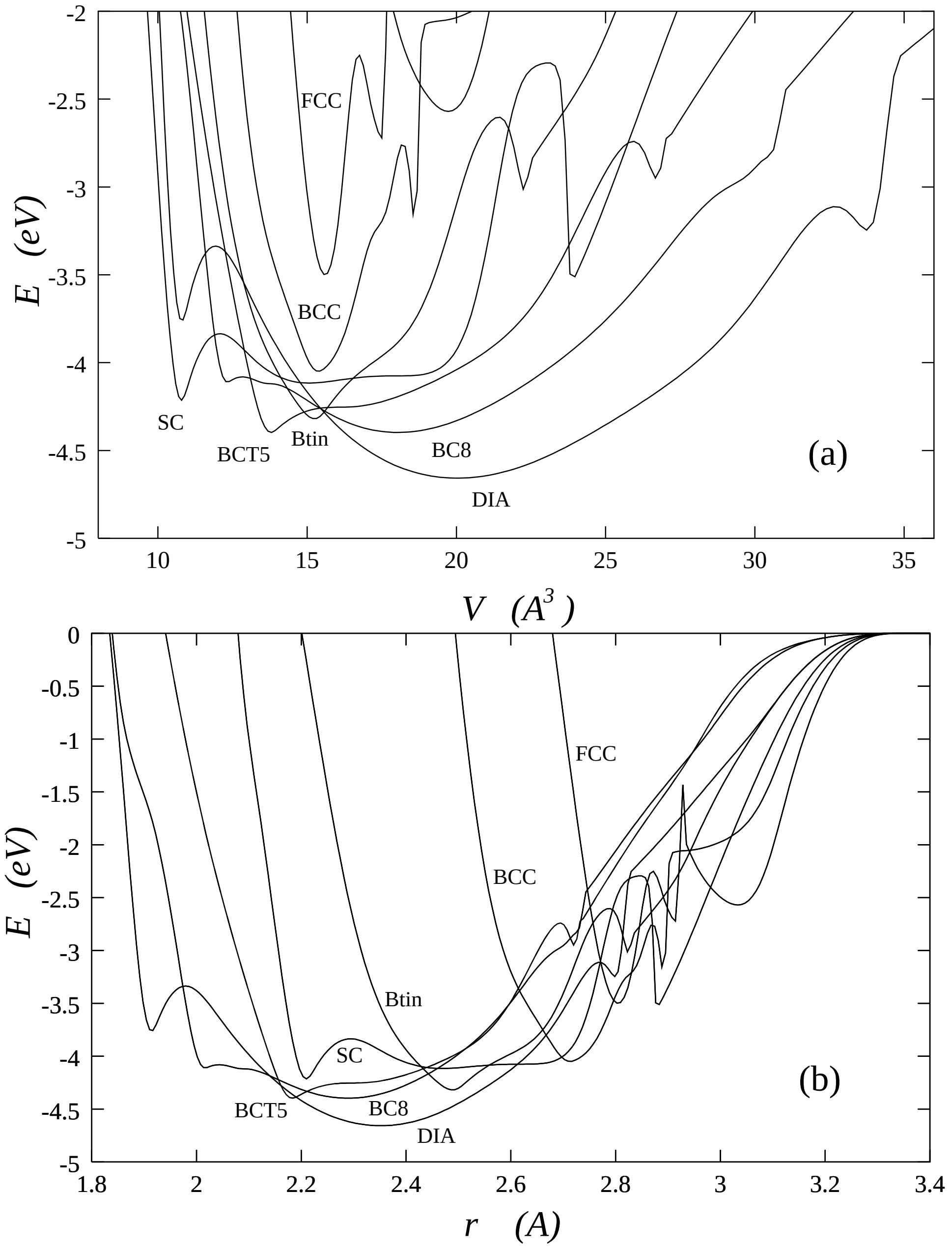

Figure 5.8: Cohesive energy curves versus volume (a) and first neighbor distance (b) computed with EDIP for various silicon crystal structures.

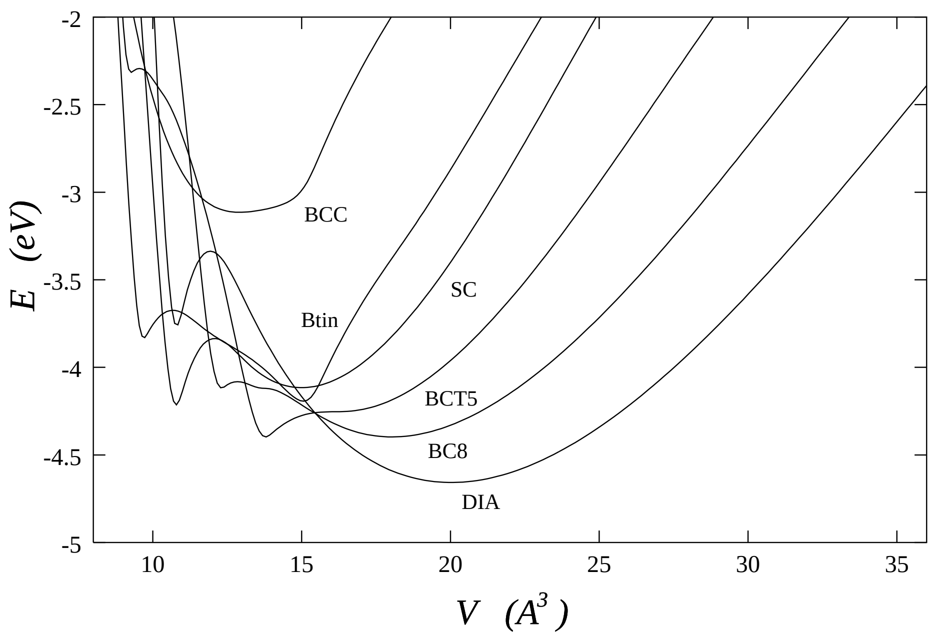

Figure 5.9: Cohesive energy curves computed with EDIP, with the coordination number artificially fixed at the correct equilibrium value for each crystal.

not removed, and thus must also have another cause.

The oscillations at small volumes are important to understand because they are the source of the most unphysical bulk properties. For example, the dips in the BCT5 and SC curves of Fig. 5.8 (a) influence pressure-induced phase transitions. With some correction possible for internal relaxation, the first high pressure phase transition of EDIP is from diamond to one of these structures, probably BCT5, and not to the experimentally observed β-tin. The primary reason for these dips is not changing coordination due to second neighbors, which occurs at a slightly smaller volume when second neighbors have $r < 3.00$ Å, as seen in Fig. 5.6 (and actually helps raise the energy by weakening pair bonds). Instead, the unphysical decreases in energy for small volumes occur where second neighbors fall into the range $3.05 < r < 3.20$ Å. This is the distance where the potential permits two-body attractions but not three-body repulsions. This phenomenon

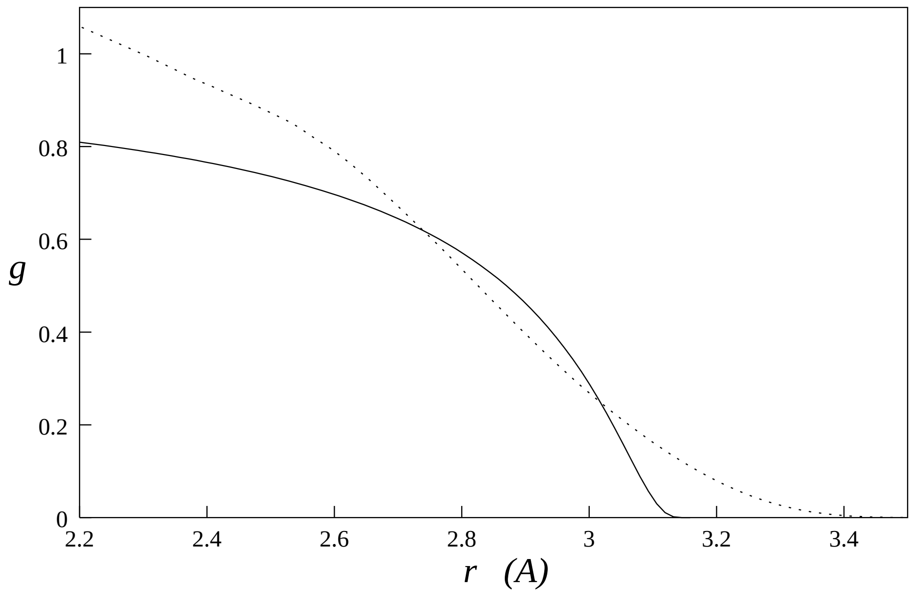

Figure 5.10: Three-body radial functions $g(r)$ for silicon from (a) the fitting of EDIP and (b) the inversion of the cohesive energy curves.

is a direct consequence of our allowing the three-body cutoff b to drift to a smaller value than the pair cutoff a in the fitting process. In general, there is an artificial and sharp decrease in the cohesive energy every time another shell of neighbors passes into the intermediate range between pair and three-body interactions. We shall see later that this is a serious problem for molecular dynamics simulations of disordered phases, but it should be possible to rectify it by slightly increasing the three-body cutoff. Refitting with $b = a$ for $g(r)$ (with coordination neighbor function the same, using the original $b < a$) is currently being explored.

The main factor leading to the sharpness of these and the coordination-induced oscillations is the sudden rise in the fitted $g(r)$ at the cutoff b. Our experience with three-body inversion clearly dictates that $g(r)$ must rise very gently near the cutoff. Fig. 5.10 compares the fitted $g(r)$ with the inverted function of Fig. 4.11, showing the relative abruptness of the fitted function. Requiring a larger value of γ (say, $\gamma > 0.5$) during fitting should remove many of these problems while preserving energies of covalently bonded structures.

5.2.4 Discussion

In summary, by fitting the EDIP functional form to a carefully chosen database of bulk crystal properties and defect formation energies, a superior potential has been produced, with a well-defined range of applicability. The potential performs remarkably well for bulk material near the equilibrium density where most bonds are covalent sp^3 hybrids. This includes isolated defects embedded in a nearly perfect diamond crystal with local relaxation and reconstruction. If the volume is increased so that bulk chemical bonds are broken, then the theory behind the functional form breaks down, and, understandably, the potential predicts unrealistic forces. At smaller volumes typical of metallic bonding, we have also observed unphysical effects, but theory suggests that the EDIP functional form may be refit to predict correct transitions and crystal stability of metallic phases. There are always going to be problems associated with neighbors passing through the cutoffs and contributing only partially to the coordination, but the hope is that they may not influence the most important structures, namely those corresponding to energy minima and saddle points, which tend to involve strong bonds in bulk defects. The only really serious problems we have found, after unusually comprehensive testing for bulk crystals and defects, are the low volume oscillations in cohesive energy curves, but these have been linked to the shorter range of many-body forces compared to pair interactions. It is possible that with minor refitting a transferable EDIP potential for bulk phases and defects can be produced, which would be a major advance over existing empirical models for silicon. In the meantime, it is interesting to ask how the current version of EDIP performs for disordered phases, like the liquid and amorphous states.

Chapter 6

Molecular Dynamics Simulation of Disordered Phases

> Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.
>
> – David L. Goodstein [132]

It is with well-deserved reservations that we now embark on a study of disordered phases of silicon with the latest version of EDIP. Although our experience with cohesive energy curves may foreshadow problems with the liquid, a metal of greater density than the solid, the theory behind EDIP certainly addresses metallic bonds and overcoordination, so we may hope for a reasonable liquid. The success of the current EDIP for defect structures and the bulk crystal indicates an exceptionally good description of sp^3 hybrid bonds, so it would seem likely that the same version would perform well for the amorphous phase, which is made up of a random network of distorted tetrahedra. These expectations can only be checked by molecular dynamics simulations, using a sufficiently large number of atoms to minimize the influence of periodic boundary conditions and long enough times to obtain accurate thermal averages. In the first section we outline some of the techniques required to perform such simulations on high performance computers.

It is customary in the literature to describe these kinds of studies as computer simulations of a real material (*e. g.* silicon), which is somewhat misleading. It is more accurate to say that one is exploring the properties of a fictitious material characterized by a particular empirical potential, which may only bear some resemblance to the real material. In this chapter, we examine disordered phases of the current EDIP material through molecular dynamics simulation, including various liquid, glassy and amorphous specimens. The purpose of these of kinds of studies is to explore the transferability of the potential and to develop large-scale simulation techniques. With such knowledge, we may hope to improve the potential and as well as our ability to perform computer experiments to the point where reliable theoretical predictions for real silicon might be possible. The chapter closes with an outlook on the future of EDIP as a transferable model for silicon condensed phases and bulk defects.

6.1 Computational Methods

The molecular dynamics simulations discussed in this chapter were performed with program written in C originally designed for simulations of inhomogeneous systems of atoms interacting via short-range forces on the Connection Machine 5, a distributed-memory, massively-parallel architecture[1]. For the purposes of developing a transferable

[1] The initial work was done at Thinking Machines Corporation (a practicum experience supported by a Computational Science Graduate Fellowship from the Office of Scientific Computing of the Department of Energy) in collaboration with B. Larson, who later ported the SW force routine to the Silicon Graphics Power Challenge. The author would also like to acknowledge discussions with N. Bernstein, E. Kaxiras, P. Tamayo and J. F. Justo that were invaluable in bringing the code to its present state.

potential (a prerequisite for meaningful large-scale simulations), smaller simulations on serial machines suffice, but the foundations of large-scale parallelism have been laid. The code has data structures and some subroutines supporting the Single-Program-Multiple-Data (SPMD) parallel programming model where each processor "owns" a subset of the particles corresponding to a particular region of physical space, and continuously keeps track of "shared" particles owned by other processors that exert forces on its particles [139, 140, 141, 142, 143]. The original goal was to efficiently load balance and optimize communication patterns on for inhomogeneous atomic systems (with large density fluctuations) where the usual method of assigning each processor an identical rectangular chunk of physical space is inefficient, but such sophistication is not necessary for simulations of bulk phases of silicon, because density fluctuations are small.

In this section, we discuss the capabilities of the code for medium-scale simulations ($N \sim 10^3 - 10^4$), paying special attention to the efficiency of force computation with EDIP. The current version of the code is optimized for various serial workstations (Sun Sparc 5 and IBM RS-6000), and also can run in serial or parallel mode on a Silicon Graphics Power Challenge with four RS-8000 processors. On RS-10000 processors, the increased sophistication for out-of-order instructions increases speed by almost a factor of two, since molecular dynamics involves significant memory indirection. We also discuss the necessary modifications for larger ($N \sim 10^5$) parallel simulations.

6.1.1 Scaling with System Size

Within the program, a "particle" is a C structure living in a global array, containing flexible attributes like its mass and position, velocity and acceleration vectors. Linear $O(N)$ scaling of the force computation with the number of particles, is achieved using Verlet neighbor lists that store the interaction topology[144]. This involves keeping an evolving list of the "neighbors" of each atom, including not only atoms which are currently within the cutoff (and exerting forces) but also atoms beyond the cutoff by less than some safety distance δr, chosen large enough that more distant atoms cannot cross the safety distance and enter the interaction range before the list is reconstructed. The overhead in memory and computational time for creating and using the Verlet list increases as the safety distance grows, but a larger safety distance allows for less frequent list updates. Thus, there is an optimal choice of δr and the update interval of m time steps (of length δt) for every simulation. For example, $\delta r = 0.03$ Å, $m = 40$ and $\delta t = 0.005$ reduced units are good choices for simulation of an EDIP liquid at $T = 2500$ K, $P = 0$.

In the current version of the code, Verlet lists are constructed using the naive $O(N^2)$ algorithm (a double loop checking every pair of atoms), which dominates the linear scaling of the (much more intensive) force computation once the system size exceeds 10^5 particles. In order to simulate larger systems, it will be necessary to introduce an $O(N)$ algorithm to build the neighbor lists. For systems with homogeneous density, a successful algorithm involves sorting particles into cells based on their physical location[140, 142, 143]. The fairly small simulations presented here, whose primary purpose is validation of the potential, do not require such sophistication, but once we are ready for large-scale production runs with a reliable model, it will be beneficial to achieve linear scaling. Considering the speed of our force routine described in the Section 6.1.3, simulations of more than 10^5 atoms should be possible on our four-processor machine.

6.1.2 Dynamics and Measurements

Time integration of Hamilton's equations of motion is accomplished with a fifth order Gear predictor-corrector algorithm[144][2]. Several statistical mechanical ensembles may be probed: Of course, the microcanonical ensemble (constant NVE) is available, in

[2] Faster, but less accurate, lower order methods with smaller memory requirements are also available: second order velocity Verlet and fourth order Gear predictor-corrector.

which the classical trajectory is followed, exactly conserving energy to within numerical error. Contact with a heat bath is simulated by rescaling of velocities to set the kinetic temperature $\mathcal{T}$ using the equipartition theorem. To minimize the artificial influence on the dynamics, rescaling is only done intermittently, typically once every 50 time steps or equivalently several times during a phonon vibrational period. In this way thermodynamic averages can be measured in the constant $NV\mathcal{T}$ ensemble can be measured, which differ from the canonical ensemble (constant NVT) averages by $O(N^{-1/2})$. Finally, the isobaric (constant NPE) and isothermal-isobaric (constant NPT) ensembles can be simulated using the Andersen piston, an extended system method in which the volume is considered an additional degree of freedom that controls isotropic expansion of the system[145][3]. The kinetic pressure is measured using the virial theorem, with the internal virial computed efficiently within the force subroutine. The program also supports simultaneous integration of tangent space trajectories for measurement of the Lyapunov spectrum using the method of Bennetin[147, 148], which can be useful in studying connections between chaotic, classical dynamics and statistical mechanics.

In order to save disk space, measurements of statistical quantities are accumulated as the program runs, thus eliminating the need to periodically store the state of the system (except once at the end for purposes of restarting the simulation). In addition to the usual thermodynamic averages (T, P, E, V) and their root-mean-square fluctuations, several properties of the EDIP potential energy function are monitored, namely the coordination number and the pair and the three-body energies, which provide information about the type of chemical bonding present in the system. Average local structure in time and space is measured with the pair correlation function and bond angle distribution. Thermal averages of these structural quantities are usually taken at

[3]This method is useful for situations in which no spontaneous spatial symmetry-breaking is expected, as in the liquid and amorphous phases. For crystal phase transitions, methods like that of Parinello and Rahman are required to allow for changes in the shape of the simulation box as well as external shear stresses[146].

500–1000 time step intervals over 10,000–50,000 steps of dynamics. For larger systems, less time is needed for good averages due to the equivalence of temporal and spatial averaging of local physics in the thermodynamic limit.

For nonequilibrium simulations, the temperature and pressure may be controlled dynamically in several ways. The external pressure is imposed with the Andersen piston, and the temperature is controlled with velocity rescaling. Each can be held constant or ramped up or down smoothly using a cosine or linear profile in time. Instead of controlling the temperature explicitly, which involves sometimes adding and sometimes removing heat (kinetic energy), it is also possible to add or remove heat at a constant rate by an appropriate choices of velocity rescalings, which permits fluctuations in kinetic temperature. Following a crystal lattice start, the persistence of crystal structure is measured with the translational order parameter,

$$\rho(\vec{k}) = \frac{1}{N}\sum_{i=1}^{N}\cos(\vec{k}\cdot\vec{r}_i), \tag{6.1}$$

where $\vec{k}$ is a reciprocal lattice vector, like $(2\pi/a)(-1,1,-1)$ for FCC or diamond, where a is the lattice constant. For solids the order parameter is of order unity, and for liquids it oscillates about zero by the usual scaling with system size, $O(N^{-1/2})$. Equilibration following a lattice start or other nonequilibrium period is achieved once the sample distribution (mean and variance) of each thermodynamic variable converges to a fixed distribution.

6.1.3 Efficient Force Computation

Force computation is the primary bottleneck in molecular dynamics, even with the simplest interaction model (a pair potential). With an efficient $O(N)$ code, force computation takes around 70% of the total time for a large scale simulation with a pair potential and can take over 90% of total time with a many-body potential. These percentages reflect a fully optimized code, which is often faster than one's naive first attempt by several orders of magnitude.

m-Loop Potentials

A many-body potential like EDIP offers a challenging computational problem. The term "many-body" (or "N-body") typically refers to a potential containing nonlinear combinations of sums over neighbors, but from a computational point of view there are varying degrees of "many-body-ness". The embedded atom potentials[101] used for metals often have the form of the EDIP pair interaction, where coordination Z_i (a pairwise sum of radial quantities) is interpreted as a background embedding electron density, which, of course, is roughly proportional to the number of neighbors. Although all neighbors contribute nonlinearly to the energy, such a potential is effectively a three-body potential as far as computation is concerned: there can be forces on a third atom k from bond (ij) due to its changing the electron density (or coordination) of the bond. Therefore, let us introduce "three-loop" as a more descriptive term for the complexity of such a potential, rather than "many-body". The Tersoff potentials are also three-loop potentials, but the three nested loops arise from explicit angular terms in the bond order: a third atom k interacts with bond (ij) based on its angle θ_{ijk} as well as its distance from the bond r_{ik}. EDIP is more sophisticated, because it is a four-loop potential. The pair interaction requires three-loops, but due to its coordination dependence, the three-body interaction requires four nested loops: a fourth atom l generates forces on triplet (ijk) by changing the coordination of the central atom i.

Nested loops greatly increase computational time; the naive algorithm for an m-loop potential is $O(N^m)$. With neighbor lists the scaling is $O(N)$, but there is a possibly large prefactor n^{m-1} multiplying the computational time, where n is the average number of neighbors per atom. Thus, the scaling of three-loop potential is $O(Nn^2)$ with neighbor lists, and naïvely, a four-loop potential like EDIP should scale like (Nn^3). However, we shall see that the increased sophistication of EDIP comes at very little computational cost. The reason is that the coordination changes only when atoms lie in the range $c < r < b$ where $df/dr \neq 0$. The number of such atoms is usually very small, often zero, in any simulation. Thus, the computational efficiency of EDIP, $O(Nn^2n_c)$, is essentially the same as a simple three-body potential like SW, where n_c is the number of neighbors with $0 < f < 1$ in the range $c < r < b$.

Neighbor Lists

The choices of book-keeping methods, loop structures and factorizations can significantly affect the overall speed. First consider the book-keeping of neighbors, the key to linear scaling. For a pair potential, each "bond" or pair of atoms is stored once in the overall neighbor list[4], and in the force subroutine each atom gets and equal and opposite force (computed once per bond) by Newton's third law. This case has been discussed by many authors, but little has been published about the computation of many-body forces. For a three-body potential like SW, it might seem reasonable to store a list of every unique triplet of atoms, but this turns out to be rather inefficient and cumbersome, for many reasons. With such an approach, neighbor lists must store atoms at twice the cutoff distance, because an interacting triplet can be as elongated as a linear chain. This means that many bond lengths stored in the neighbor list will not make any force contribution. The size of the list is also increased compared to a simple Verlet pair list. This approach also poses major problems for SPMD parallel programming. As stated above, the most successful approach for short-range atomistic dynamics is to assign atoms to processors, with communication only required for interactions between atoms owned by different processors. The problem with a triplet-neighbor list is that it is more difficult to assign forces to the correct atoms and structure communication because there are multiple possibilities for triplet ownership, like own-own-own, own-own-shared, own-shared-own, shared-own-own, own-shared-shared, ... These problems grow quickly with the order of the interaction. The best solution for many-body forces

[4]These comments also apply to the cell-based algorithms, in which case the role of the Verlet list is played by the cell interaction graph[140, 142, 143].

is to store a list of *all* neighbors within the cutoff range of *each* atom. This causes the force routine to visit each interacting pair twice, each triplet three times, ..., but it turns out to be far more efficient. This is particularly so in the case of EDIP, because forces are not symmetric and depend on the environment (coordination) of only one atom at a time in each pair or triplet.

Force Algorithm

The design of an optimal loop structure and factorization of the derivatives is much more subtle than the choice of book-keeping method. Since the pair and three-body interactions depend on the coordination, a prepass over pairs of atoms is required. Because separation distances are checked in computing the coordination, it makes sense to precompute all radial quantities in the initial pair loop, which are stored in small temporary arrays (reused for each atom). The subsequent loops for many-body forces are made over these temporary arrays, thus eliminating the need to unnecessarily check if neighbors lie within the cutoff. In factoring the derivatives, the guiding principle is to push every computation into the outermost possible loop. With the optimal decomposition, the innermost (fourth) loop for three-body coordination forces only involves a single addition operation, performed only for previously identified neighbors in the range $c < r < b$. All forces from changing coordination are applied in an extra pair loop after the pair and three-body force loops. Pseudo-code for the algorithm is given in Fig. 6.1. In summary, the optimal algorithm consists of a sequence of four passes over the neighbors of each atom: (1) a pair loop to compute the coordination Z and store radial quantities sorted by interaction distance; (2) a pair loop to compute $V_2(r, Z)$ with a nested three-body loop to accumulate terms for coordination forces; (3) two nested three-body loops for $V_3(r_1, r_2, l_{12}, Z)$ with a nested four-body loop for coordination terms; (4) a pair loop to apply coordination forces. We shall see that this algorithm handles the sophisticated 4-loop EDIP forces with remarkable efficiency.

```
for each (own) atom i
   Z = 0
   for each neighbor (own or shared) j of i
      if r_ij < a
         store radial factors in V_2 and dV_2/dr (index k, nk[k] = j )
         if r_ij < b
            store g, dg/dr and r⃗_ij (index m, nm[m] = j)
            if r_ij < c
               Z = Z + 1
            else
               Z = Z + f
               store df/dr and r⃗_ij (index n, nn[n] = j)
   compute p(Z) and dp/dZ
   for each n, zero a scratch array sz[n] = 0
   for each k
      accumulate energy V_2
      if i < nk[k], apply forces dV_2/dr to i and nk[k]
      compute dV_2/dZ
      for each n
         sz[n] = sz[n] + dV_2/dZ
   compute w(Z), dw/dZ, τ(Z) and dτ/dZ
   for each m_1
      for each m_2
         compute cosine l_12 = r⃗_i1 · r⃗_i2
         accumulate energy V_3
         apply forces (dg_1/dr)g_2h_12 to i and nm[m_1]
         apply forces g_1(dg_2/dr)h_12 to i and nm[m_2]
         apply forces g_1g_2(dh_12/dl) to i, nm[m_1] and nm[m_2]
         compute dV_3/dZ = g_1g_2dh/dZ
         for each n
            sz[n] = sz[n] + dV_3/dZ
   for each n
      apply forces sz[n] · (df/dr) to i and nn[n]
```

Figure 6.1: Outline of an efficient algorithm to compute EDIP many-body environment-dependent forces. Indentation specifies the scope of a loop or conditional statement. Interpret "apply forces" throughout as "apply equal and opposite forces using Newton's third law". "Own" and "shared" refer to the SPMD parallel programming model.

6.1.4 Benchmarks

In order to test the performance of our code, consider a variety of representative simulations using the SW potential on a single 90 MHz Silicon Graphics RS-8000 processor with a 4 MB cache. The results are displayed in Table 6.1. For small simulations involving a few thousand atoms or less, the code scales like $O(N)$ on a serial machine, indicating dominance of the total time by the $O(N)$ force computation. The benchmarks around 50 μs per atom per time step show that the dynamics of a few thousand atoms can be followed for about a million time steps in just over a day of CPU time on a typical workstation. In physical terms, this means that we can simulate a $35 \times 35 \times 35$ Å^3 chunk of silicon for over 50 ns without needing a supercomputer. With this capability, we can easily investigate a number of interesting systems of experimental relevance, such as solid phase epitaxial growth, radiation damage and plastic deformation. The factor limiting the feasibility of such simulations, however, is not computational expense, but rather transferability of the empirical potential (which is insufficient with the SW potential for the aforementioned applications).

Much larger simulations are also possible with our force routine, but some work is needed to prepare the rest of the code for large-scale parallelism. The force subroutine possesses reasonable parallel efficiency once the number of atoms exceeds a few thousand, meaning that the overall scaling is $O(N/P)$, where P is the number of processors, for small-scale, shared-memory parallelism. Unfortunately, once the system gets that large, the poor $O(N^2)$ scaling of the serial neighbor list algorithm begins to dominate, diminishing any advantage gained by the efficient, parallel force routine. Once N exceeds 10^4, the simulation speed is already reduced by a factor of two, over $N = 10^3$, and speed decreases as $1/N^2$ for larger systems. These problems can be eliminated by implementing an $O(N)$ cell-based neighbor list algorithm, which would make possible simulations of up to 10^5 atoms on the Power Challenge.

Since EDIP is considerably more complex than the SW potential, and certainly more

Phase	N	P	Total Time	Force Time	Forces	Lists	Integration
Solid	216	1	56	53	94%	2%	3%
Solid	1,728	1	62	52	84%	13%	3%
Solid	1,728	4	55	31	56%	33%	11%
Solid	12,288	4	138	28	20%	75%	5%
Liquid	1,728	1	125	166	75%	21%	4%
Liquid	1,728	4	83	59	70%	22%	8%

Table 6.1: Timing analysis of our molecular dynamics program. The effects of thermodynamic phase (equilibrium liquid at $T = 2500$ K and solid at $T = 300$ K), system size (N particles) and small-scale parallelism ($P = 1$ or $P = 4$ processors) are demonstrated for systems in the microcanonical ensemble interacting via the SW potential on a Silicon Graphics RS-8000 Power Challenge. The total simulation time and force time are in μs per particle per time step, and percentages of the total time are given for force computation, neighbor list construction ($m = 100$ for solid, $m = 50$ for liquid, $\delta r = 0.03$ Å) and time integration (velocity Verlet scheme).

Phase	LJ	SW	IK	EDIP
Solid	49	55	46	54
Liquid	43	135	130	98

Table 6.2: Comparison of the speed of force computation with the Lennard-Jones (LJ), Stillinger-Weber (SW), Ismail-Kaxiras (IK) and EDIP potentials for typical solids and liquids (in μs per atom per time step on a Silicon Graphics RS-8000 processor).

than a pair potential like Lennard-Jones (LJ), we may worry that some of benefits of large-scale simulation just described might be lost with EDIP, but that is not the case at all. As shown in Table 6.2, the optimized EDIP force routine described earlier outperforms both the SW and IK potentials in a fair comparison of equivalent simulations. The solids mentioned in the table (for SW, IK and EDIP) are diamond structures at $T = 300$ K (velocity rescaling) and $P = 0$ (Andersen piston), and the liquids are at $T = 2500$ K and $P = 0$. There are two reasons for the impressive speed. The first is that the many-body coordination forces do not arise often and are handled very efficiently when they do. The second is that the fitted EDIP has a smaller cutoff (3.45 Å) than SW (3.77 Å) and IK (3.73 Å). The result is that many fewer interactions need be considered, especially in the liquid, since the SW and IK cutoffs lie just short of the second neighbor distance (3.84 Å) in the solid. With a many-loop potential, speed is very sensitive to the number of neighbors.

The effect of the short cutoff of EDIP is seen more dramatically in comparison with the LJ potential. Of course a 2-loop pair potential is much faster to compute that a 4-loop many-body potential, but we are lucky that, although silicon bonding is more complex, it involves many fewer neighbors than the Van-der-Waals bonding described by the LJ potential. For a fair comparison with the silicon potentials, we must consider representative simulations of noble elements with the LJ potential. The LJ solid mentioned in Table 6.2 is argon in the FCC structure, using the commonly used cutoff of 2.5σ for the potential at room temperature and zero pressure. The liquid is in equilibrium just above the melting point ($T^* = 1, P^* = 0$). Remarkably, in such a fair comparison of typical simulations, EDIP competes with LJ in computational speed. For the solids, EDIP is only slightly slower than LJ (as are SW and IK), and for liquids it is slower by a factor of two (while SW and IK are three-times slower). In conclusion, the EDIP functional form offers greatly increased sophistication without paying any price in computational speed. In applications for which the model is reliably validated (*e.g.*

dislocations), very large and realistic simulations of silicon are possible. For example, if our EDIP force subroutine were incorporated into an existing massively parallel code like SPaSM [142], then simulations involving 10^8 atoms would be possible.

6.2 The Liquid Phase

Using the computational machinery of the previous section, it is straightforward to study disordered phases of EDIP. In the following sections we discuss computer experiments on specimens of EDIP material to test whether current fitted potential behaves anything like real silicon when melted and quenched. These phases are quite far from the fitting database, but lie within the theoretically predicted range of validity of the functional form. We begin with the liquid phase, obtained by melting the solid.

6.2.1 Crystal Melting

Experimentally, silicon melts at $T_m = 1685$ K, undergoing a phase transition with a latent heat of 50.7 kJ/mol from a tetrahedral, semiconducting solid to a metallic liquid with a 10% smaller volume and just over six-fold coordination [149, 150]. The detailed structure of the liquid, revealed by experiments and *ab initio* calculations [151], is described in the next section, and here we investigate thermodynamic properties associated with the phase transition. Although the structure of liquid silicon is rather difficult to predict with empirical potentials, most existing models have failed to predict even the thermodynamic properties. The Tersoff potentials (T2 and T3) predict melting temperatures almost twice the experimental value[5]. The BH potential also overpredicts the melting temperature at 3000K. The only models that describe the melting transition are SW and MFF, but it is these are also the only ones that were explicitly fit to the melting temperature. The density increase from solid to liquid is underestimated by

[5]This can be improved by changing the cutoff distance [40], but it is not clear that other important properties can be preserved.

SW by a factor of two [135]. In their original paper, SW report a melting temperature of 2013 K [13], but subsequent studies find values in the range 1665–1750 K (although if the SW potential is scaled to set the correct bulk cohesive energy, then the melting temperature should scale into the range 1775–1865 K [19]). The discrepancies among the calculated melting temperatures depend partly on varied simulation techniques, but mainly on available computational resources, which restricted early studies to rather small systems (64–216 atoms) and short times (0.1–10 ps). The MFF potential, which has not been thoroughly tested overall, is reported to predict a melting temperature of 2050 K (although their simulations of a 64 atom periodic cell are probably not very accurate) [55].

In our study of the melting transition with EDIP, we are able to achieve greater accuracy using recent advances in high performance computation, as described in the previous section. Following the method of Luedtke and Landman [134], we melt the crystal at zero pressure by slowly adding heat at a constant rate. In this way we can determine the heat capacity of the liquid and solid states and the latent heat of the phase transition, as well as the melting temperature. Starting with perfect, 1728-atom diamond crystal at $T = 300$ K, we add 1.1 eV/atom over 16.5 ns (215,000 time steps) of dynamics. The heat flux of 38,600 eV/atom·sec is, of course, much faster than experiment, but it is three orders of magnitude slower than in early studies (and the system is ten times larger). We could achieve even slower rates and larger systems on parallel supercomputers, but the discontinuities associated with the phase transition are seen quite clearly with these modest simulations, which can run overnight on a serial Silicon Graphics RS-8000 processor. The first order transition takes around 6.6 ns with these parameters, including superheating of the solid, melting, and recovery from supercooling of the new liquid.

The results of the simulation are shown in Fig 6.2. Because heat is added at a constant rate, the total energy is proportional to time (from the first law of thermodynamics,

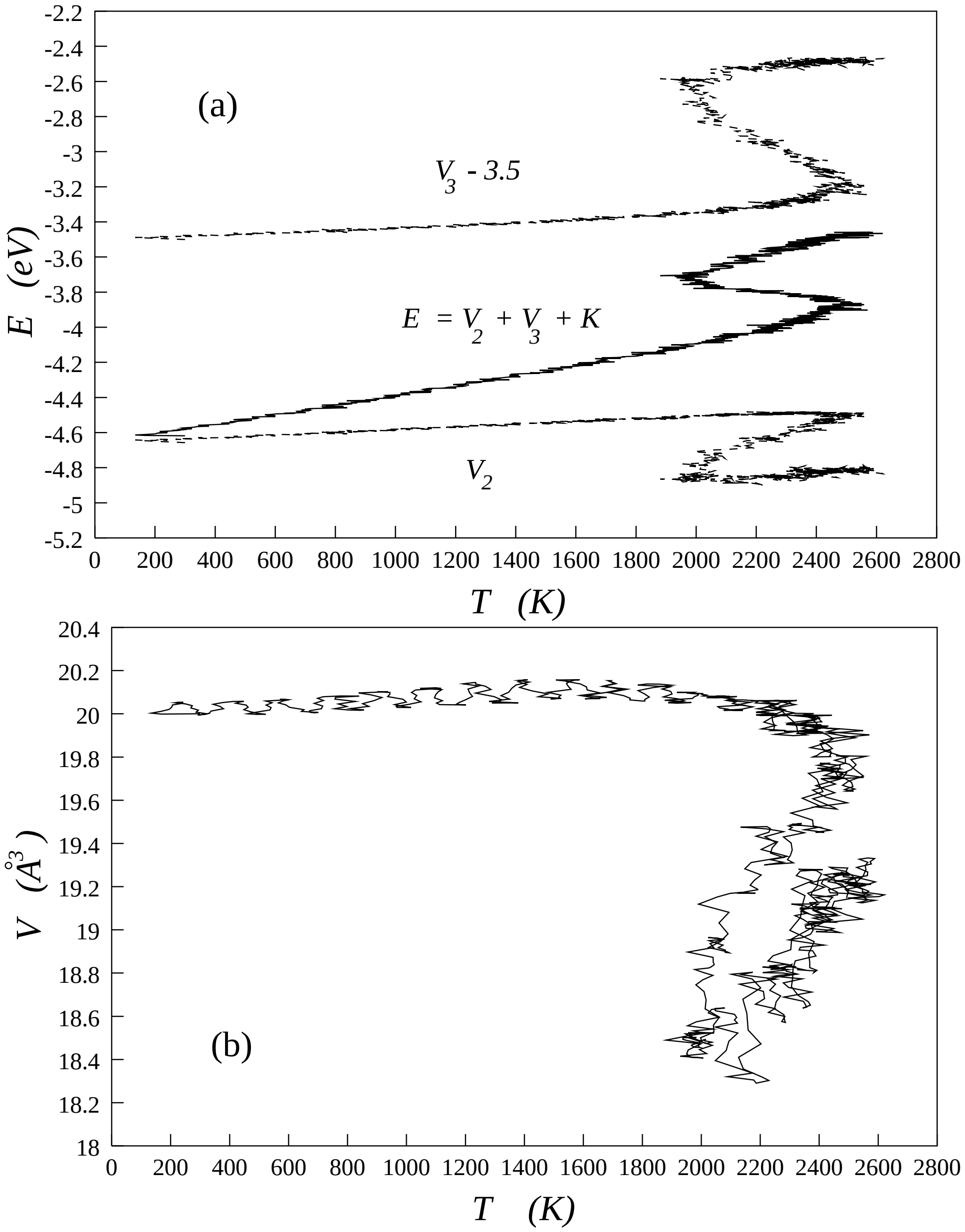

Figure 6.2: Melting of a 1728-atom EDIP solid with a constant heat flux of 38.6 eV/atom·ns. The total (E), pair (V_2) and three-body (V_3) energies as a function of temperature are shown in (a), and the volume per atom versus temperature is shown in (b).

$dE = dQ$ since $P = 0$), but the temperature fluctuates. In a unique phase (solid or liquid), the temperature rises at a roughly constant rate, but during the phase transition the temperature temporarily lowers as the potential energy rises. In the EDIP model, the pair energy actually reduces from liquid to solid (due to the increasing coordination), but the three-body energy increases to overcompensate. The heat capacity (slope of the E versus T curve) has a small temperature dependence away from the transition, with $C_p = 25$ kJ/mol·K for a solid at $T = 1000$ K, in good agreement with experiment (like the other potentials [126]). The heat capacity of the liquid $C_p = 36$kJ/mol·K at $T = 2500$ K is quite close to the experimental value 31 kJ/mol·K. (MFF predicts 28 kJ/mol·K [55].)

The unphysical temperature variation of the total energy during the phase transition reflects the finite system size and fast dynamics. The interfacial liquid-solid tension is not neglible for small nuclei, and, because there less likelihood of a nucleation event in a small, perfect crystal during a short time, the solid is easily superheated. The volume variation with temperature during the transition in Fig 6.2 (b) shows collapse into a supercooled liquid state of higher density and (large coefficient of volume expansion) before recovering to the equilibrium liquid under subsequent heating. Clearly the system is not large enough (or the melting slow enough) to allow coexistence of the two phases in equilibrium. The density increase from solid to liquid is underpredicted by EDIP at 4.4%, compared with 5.5% for SW. Coincidentally the supercooled liquid that appears during our phase transition has nearly the correct density.

From our bulk simulation, the melting temperature is between 2000 and 2500 K. It is tempting to make an equal area construction and conclude that the melting temperature is 2230 K, but such analysis is not valid in this data. A more accurate determination of the melting temperature can be made by introducing two (100) surfaces into the system by switching off periodic boundary conditions in one direction, and repeating the same melting procedure. These large defects prevent the need for small liquid nuclei

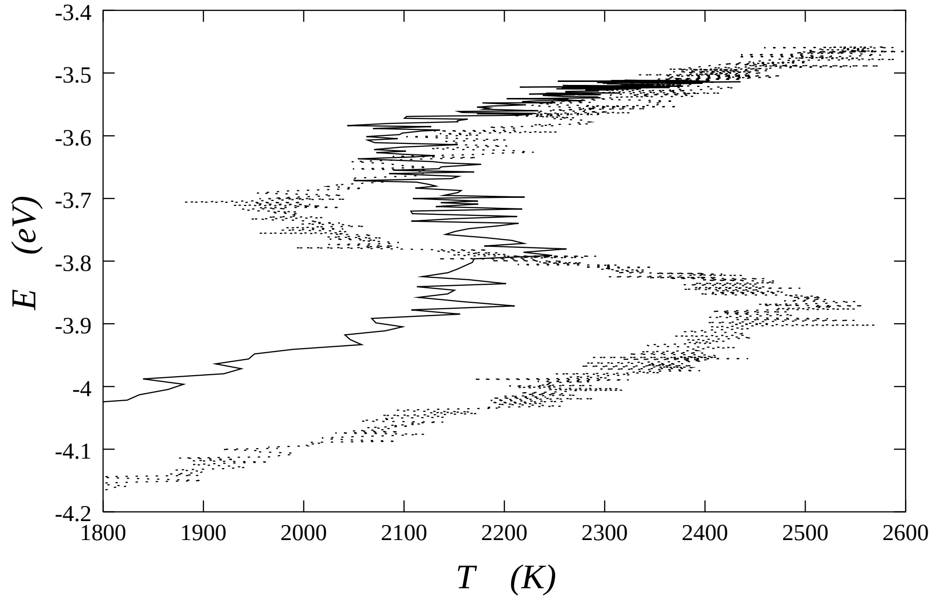

Figure 6.3: The melting transition (total energy versus temperature at constant pressure) under a constant heat flux for 1728-atom EDIP solids with (solid line) and without (dotted line) free surfaces.

growing in the bulk crystal to overcome interfacial tension, and provide numerous seeds of disorder as the surface premelts. In Fig. 6.3, we see that a fairly sharp phase transition is achieved, within the range suggested by the periodic bulk simulation, at a temperature of $T_m = 2150 \pm 50$ K. This result is in fairly good agreement with experiment, rivaled only by SW and MFF.

By measuring energy differences at the melting temperature in Figs. 6.2 and 6.3, we can obtain the latent heat of melting. EDIP predicts a latent heat of $\Delta E = 42.5$ kJ/mol (0.44 eV/atom), which is considerably closer to experiment (50.7 kJ/mol) than the SW (31.4 kJ/mol) [134] and MFF (30 kJ/mol) [55] values. The EDIP latent heat can be broken down into separate contributions of -0.38 eV/atom from the pair energy and 0.83 eV/atom for the three-body energy.

In summary, EDIP does rather well in predicting the thermodynamic properties of the melting transition. Its performance for the melting temperature, latent heat, heat capacities and volume change is comparable to SW and MFF, which were both explicitly fit to the transition, and EDIP outperforms other potentials that were not fit to the liquid. However, thermodynamic properties are only rough indicators of the realism of the model, and to better understand the physical relevance of the EDIP liquid we must look closely at its structure.

6.2.2 Liquid Structure

In order to study structural properties, the liquid is created as follows. The structure generated by this method is quite similar to that obtained by melting, but is ensured to be better thermalized, with no remnants of crystalline order. First, a perfect diamond lattice of 1728 atoms is given random initial velocities (uniform distribution for each component), rescaled to set a kinetic temperature of 5000 K. This is a very unphysical situation, an extremely super-heated solid, but it serves well to generate a random mixture of atoms with reasonable separations at nearly the right density. The system violently raises its entropy (and thus lowers its free energy) by melting, and the order parameter drops to zero in less than 50 ps. The temperature is maintained at 5000 K for 0.3 ns to thermalize the velocity distribution and to equilibrate the volume at constant pressure. The temperature is then gently ramped down to 2500 K over a period of 2.5 ns. Finally, the system is allowed to equilibrate at $T = 2500$ K and $P = 0$ for another 1.5 ns.

The average local structure of the EDIP liquid is shown Fig. 6.4. The pair correlation function $g(r)$ has a sharp peak that nearly overlaps with the first neighbor peak of the diamond solid. The first maximum is at $r = 2.38$ Å. Integrating up to the first minimum at $r = 2.84$ Å, we find that the first peak contains 4.24 neighbors. Thus, EDIP melts into a primarily covalent liquid, with four distorted sp^3 hybrid bonds per atom. In contrast, the experimentally observed liquid is metallic [149] with a coordination of 6.4 [150]. The *ab initio* liquid has coordination 6.5 up to a distance 3.10 Å [151], and a pair correlation function $g(r)$ quite similar to the SW result, shown in Fig. 6.4, although SW technically overestimates the coordination, as described below.

Although the *ab initio* liquid is metallic and overcoordinated, *ab initio* dynamical studies reveal that tetrahedral fluctuations do exist in the liquid, which helps it find the amorphous network upon cooling [151]. An analysis of *ab initio* charge densities shows that covalent bonds (charge transfer from atomic orbitals to bonding orbitals) always form between atoms closer than $r_c \approx 2.5$ Å, and larger separations break these bonds, leaving a weak metallic interaction. The EDIP liquid contains a similar mixture of covalent and metallic bonding, but the percentage of covalent bonds is overestimated. The *ab initio* liquid has roughly two first neighbors in the covalent range, $r < r_c$, while the EDIP liquid has four. The other four first neighbors in the *ab initio* liquid are in the range of metallic bonds.

EDIP also has metallic near-neighbors, but they are concentrated in a narrow, anomalous peak of $g(r)$ with a maximum at $r = 3.11$ Å in between the first and second

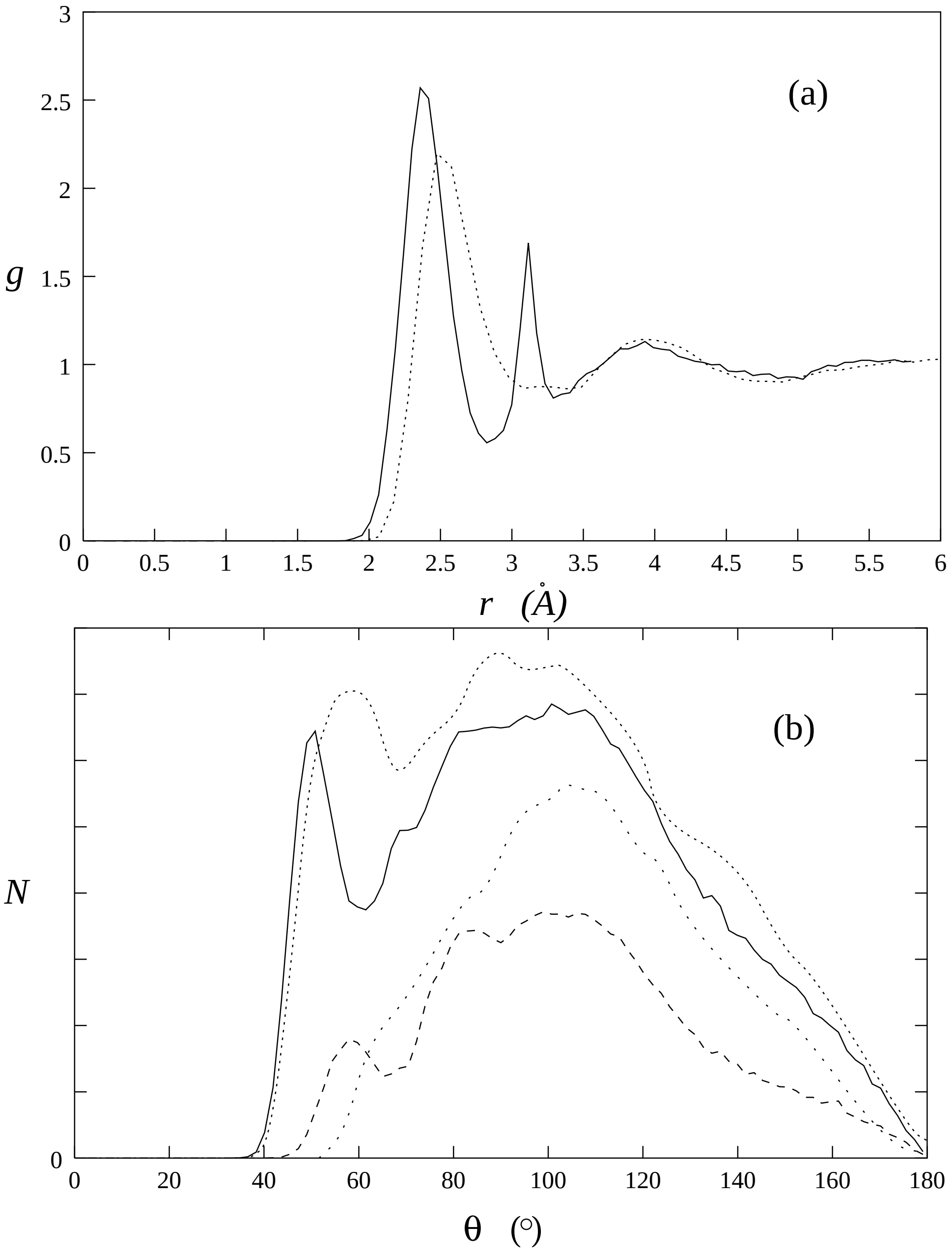

Figure 6.4: Structure of the EDIP liquid. The pair correlation (solid line) is shown in (a) and compared with the SW liquid (dotted line), which is close to *ab initio* [151]. The bond angle distribution is shown in (b) for coordination neighbors with $r < 3.31$ Å (solid line) and also $r < 2.84$ Å (dashed line), and compared with the *ab initio* distribution for $r < 3.10$ Å (dotted line) and $r < 2.50$ Å (widely spaced dotted line) [151].

neighbors in the diamond lattice. The second peak contains roughly one more neighbor. To be precise, integration up to the second minimum at $r = 3.31$ Å yields a total coordination of 5.02. Recall that we have encountered the distance $r = 3.11$ Å before; the unphysical dips in all the EDIP cohesive energy curves occur when neighbors pass through this distance and gain pair energy without paying any penalty in three-body energy due to the disparate cutoffs $b < a$. Clearly the same effect is seen in the liquid. Nevertheless, in spite of the unphysical splitting of the first neighbor peak, EDIP may be the first potential to predict a liquid with a clear mixture of covalent and metallic bonds, albeit in the wrong proportions, as evidenced by three-body correlations.

Our assessment of the EDIP liquid is refined by consideration of the bond angle distribution, shown in Fig. 6.4 (b). The atoms from first neighbor peak of $g(r)$ (dashed line) have a broad maximum around the tetrahedral angle, consistent with the sp^3 character of these bonds suggested by $g(r)$. The *ab initio* bond angle distribution (widely spaced dotted line) for atoms in the covalent range $r < r_c$ has a similar distribution. The overall *ab initio* bond angle distribution (dotted line) is broadly peaked at 90°, where EDIP first neighbors also have a small peak. If we include atoms from the anomalous second peak of $g(r)$ in the bond angle distribution (solid line), then a sharp peak develops at 50° along with a bulge at large angles, much like the *ab initio* distribution.

The structure of the EDIP liquid can be understood in terms of a simple model of the atomic arrangements. The dominant structure is a distorted tetrahedron with bond lengths close to the covalent distance 2.35 Å (first split peak of $g(r)$) and one extra atom at a metallic distance of 3.1 Å (second split peak) centered along a tetrahedral edge. The extra atom then lies at the center of a neighboring tetrahedron, with the atom at the center of the first tetrahedron playing the role of the extra atom for the second. So, the model is that of two distorted tetrahedra joined at a common edge. The four-atom ring created between the two tetrahedra is shown in Fig. 6.5. The angle formed between first neighbors at the edge is 75°, which is consistent with the skewing of the tetrahedral

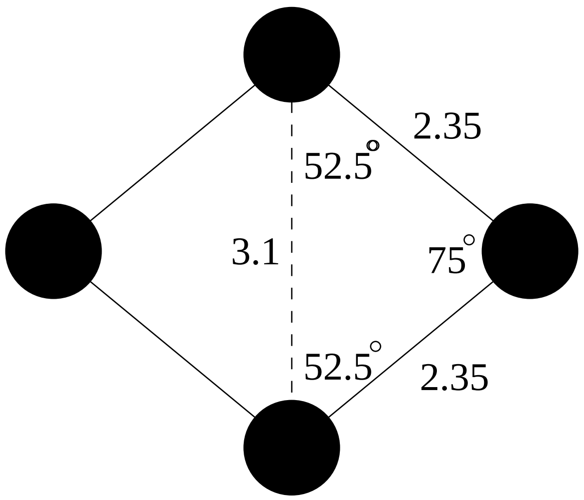

Figure 6.5: A typical structure in the EDIP liquid, which may also appear in the *ab initio* liquid, possessing a mixture of covalent and metallic bonds.

peak in the bond angle distribution toward the range 80–90° in Fig. 6.4 (b). The angle between the extra atom and the edge atom is 52.5° which explains the sharp peak just above 50° in the bond angles between the two split peaks in the $g(r)$. Similarities with the *ab initio* bond angle distribution suggest that this local structure may appear in the real silicon liquid as well.

Another sensitive measure of liquid structure is the distribution of local coordinations, defined as the number of neighbors closer than the first minimum (beyond the first peak) of $g(r)$. The numerical results are given in Table 6.3 and plotted in Fig. 6.6. The *ab initio* distribution resembles a bell curve of half-width 1.5 centered at 6.5. For a fair comparison with EDIP, we must include in the coordination neighbors from the first two peaks with $r < 3.31$ Å, which clearly represent a splitting of the first peak (into covalent and metallic subshells as discussed above). The EDIP distribution is remarkably close to *ab initio* across the entire range of local coordinations, a feat which has not been reported for any other potential. The table also gives coordinations for neighbors in the inner subshell $r < 2.83$ Å, which is peaked at 4, in agreement with the

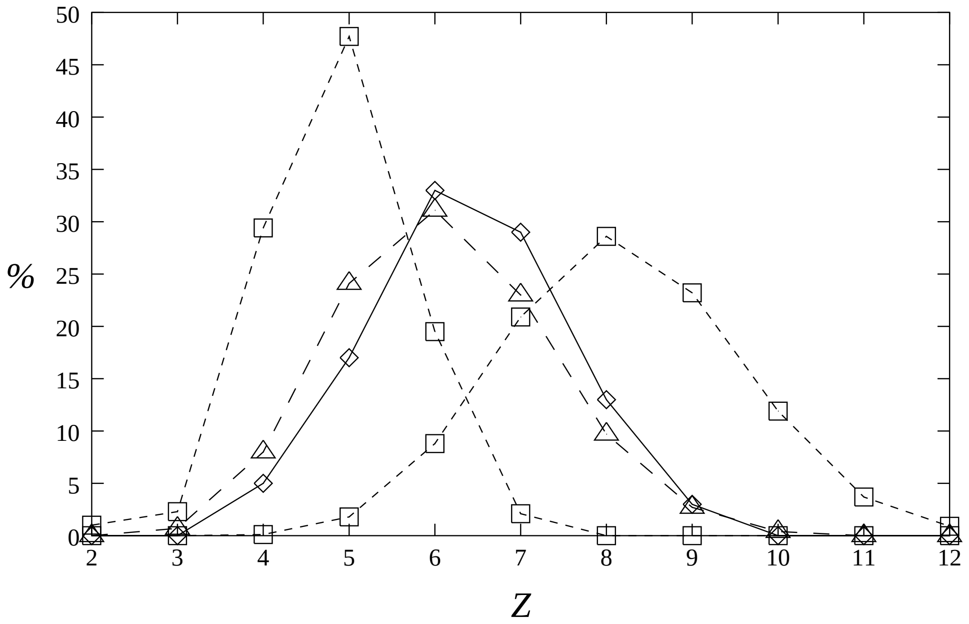

Figure 6.6: The distribution of atomic coordinations in the *ab initio* (diamonds), EDIP (triangles) and SW (squares) liquids. The higher Z SW curve is for the true coordination cutoff 3.43 Å, and the other is for a shorter cutoff of 3.0 Å [134].

Z	LDA	EDIP	SW	SW 1	EDIP 1	EDIP Z
2	0	0	0	1	0	0
3	0	1	0	2	16	5
4	5	8	0	29	48	36
5	17	24	2	48	30	45
6	33	31	9	20	5	13
7	29	23	21	2	0	1
8	13	10	29	0	0	0
9	3	3	23	0	0	0
10	0	0	12	0	0	0
11	0	0	4	0	0	0
12	0	0	1	0	0	0

Table 6.3: Distribution of local coordinations (in %) for the *ab initio* (LDA) [151], SW ($r < 3.43$ Å) and EDIP ($r < 3.31$ Å) liquids at 2000 K. For EDIP, separate statistics are given for neighbors under the inner split first-neighbor peak of $g(r)$ (EDIP 1), and for SW we also show published data [134] (SW 1) presumably using a cutoff around 3.0 Å. The distribution of effective coordination numbers (EDIP Z) is also given, rounded to the nearest integer.

ab initio analysis of atoms inside the covalent cutoff r_c [151]. Once again, we see the coexistence of covalent and metallic bonds in both the EDIP and *ab initio* liquids.

The best known description before EDIP is by SW, but as seen in Fig. 6.6 it is not very accurate. However, the comparison depends on the definition of coordination. There is some discrepancy in the literature regarding the coordination of the SW liquid, which we may attribute to the fairly flat shoulder in $g(r)$ between the first and second neighbor peaks. The inner edge of the shoulder is at 3.0 Å, and our calculations suggest that Broughton and Li must have used this distance in getting their coordination statistics (SW 1 column in Table 6.3), yielding an average coordination of 4.97 [133]. Their (inner) coordination distribution for SW is sharply peaked at 5, a fact attributed by Kaxiras and Boyer to the low energy of the BCT5 phase predicted by SW [106]. The first minimum of $g(r)$, however, is at the other end of the shoulder at 3.43 Å, as seen clearly in the data of Broughton and Li [133]. This is the more appropriate cutoff for the coordination, yielding an average of 8.16 with our data, which is consistent with other studies reporting values of 7.7 [135] and 8.0 [13]. The full distribution for the SW liquid greatly overestimates coordinations, as seen in the figure.

Table 6.3 also shows the distribution of effective coordination numbers in the EDIP liquid, defined in Chapter 5.1. The true coordinations are severely underestimated by the EDIP coordination number ($\overline{Z} = 4.23$). We have already pointed out the underestimation of coordination when studying cohesive energy curves, but here the situation is worse. Incorrect effective coordination numbers undermine the theoretical basis of the potential, and might lead to problems. This has occured because the fitting database of bulk defect structures has very few neighbors in the range $3.0 < r < 3.2$ Å, which are plentiful in the liquid.

6.2.3 Discussion

In summary, the EDIP liquid has a number of reasonable features, even though the pair correlation function is clearly artificial. In particular, the bond angle distributions for the covalent and metallic bonds of the EDIP liquid are quite similar to their *ab initio* counterparts, a rather subtle effect. In contrast, the SW liquid bond angle distribution does not contain angles with metallic character at any distance, having simply a broad peak at the tetrahedral angle for all first neighbors [133, 134]. EDIP also captures the statistics of local coordinations much better than SW. It is often said that SW offers an excellent a description of the liquid state, indeed the best of any of the popular potentials. While SW does predict an accurate melting temperature and pair correlation function, it is worth emphasizing that upon closer inspection, the local structure of the SW liquid is wrong, as indicated by the coordination statistics and bond angle distribution. The behavior of EDIP is encouraging because it was not fit to the liquid (as was SW) or any other overcoordinated structure. The performance of EDIP for the liquid is purely an extrapolation from crystalline bonding states, made with the theoretically motivated coordination dependence of the functional form.

So, perhaps the EDIP liquid has some new and relevant physics not contained in existing models, which blindly try to enforce tetrahedral order in every situation. Although it has numerous flaws, EDIP is at least capable of modeling multiple bonding states, which opens the possibility of a new degree of transferability. It is interesting to note that if the first two peaks of $g(r)$ were merged at the weighted average distance of $r = 2.5$ Å (coincidentally, the location of the *ab initio* peak), then the structure of the EDIP liquid would be fairly realistic. Perhaps, if the abrupt rise in the three-body radial function we noted earlier were smoothed by increasing the fitting parameter γ, the liquid structure might be improved.

6.3 Amorphous Phases

Most potentials predict a quench from the liquid into a glassy phase characterized by frozen-in liquid structure. Real (and *ab initio*) silicon, on the other hand, quenches into an amorphous phase consisting of a random tetrahedral network of distorted sp^3 covalent bonds. It has been proposed that the seeds of crystalline order are created from the fluctuating covalent bonds within the first neighbor shell of the liquid described above ($r < r_c$) [151]. Using empirical potentials like SW [133, 134, 135] and BH [152], it has always been necessary to make artificial changes, usually strengthening the three-body interaction, in order to guide the system into an amorphous phase when quenching from the liquid. In these simple models, there is only one physical principle at work: the balance between the three-body forces that favor tetrahedral angles and thermal fluctuations that tend to break them. The angular preference must be weak enough to be overcome by thermal vibrations at the melting temperature, but apparently this level of angular force is not enough to overcome the free energy barrier to recover the tetrahedral structure upon cooling. The amorphous is regained by simply adjusting this balance, controlled by a single parameter, the strength of the three-body interaction.

The EDIP interaction model is more complicated, and hence has a more complex phase diagram. This make it possible to describe more subtle physical properties like the bond angle and coordination statistics in the liquid and the structure of crystalline defects, but it opens more ways for the potential to bypass tetrahedral bonding at low temperatures. Unlike other potentials, EDIP quenches into an amorphous phase that is qualitatively very different from the liquid, containing quasi-crystalline short-range order. Unfortunately, the local order is not tetrahedral.

6.3.1 The Quenched Liquid

The quench is performed by gently ramping the temperature of a well-equilibrated 1728-atom liquid down from 2500 K to 300 K in 50 ps at zero pressure, resulting in an amorphous structure we may call a-EDIP-I. The liquid-amorphous transition is second order (or perhaps even a sequence of second order transitions), with a discontinuity in heat capacity $(dE/dT)_p$ and coefficient of volume expansion $(dV/dT)_p$ around $T_a = 670 \pm 30$ K. The energy at $T = 300$ K and $P = 0$ of the EDIP amorphous phase is -4.497 eV/atom, only 0.16 eV/atom higher than the ground state crystal and lower than any other crystal phase from Chapter 5.2.3. The EDIP amorphous volume is quite low, only 16.46 Å^3/atom. The corresponding density, 0.0608 Å^{-3}, is 22% larger than the equilibrium diamond solid, while experimentally amorphous silicon has a 1% *smaller* density than the crystal. This surprising result foreshadows highly unphysical properties of the EDIP amorphous phase, which we may expect to be related to overcoordination and metallicity.

With the exception of a tall and narrow second peak[6], the pair correlation function for the EDIP amorphous is fairly close to the *ab initio* amorphous [153] as shown in Fig. 6.7. The first peaks are almost identical, with a maximum just above the diamond crystal bond length of 2.35 Å, but the EDIP peak is not quite as sharp and contains a few too many neighbors, 4.32 instead of 4.00. The non-glassy nature of the EDIP amorphous phase is seen in the vanishing pair correlation in the regions between the first three peaks. As in the liquid, there is an anomalous second peak at $r = 3.12$ Å, with the next minimum at $r = 3.26$ Å defining the true coordination distance. The second peak is much larger than in the liquid, containing around four more neighbors for a total coordination of 8.23. The eight-fold coordination may suggest a BCC-like random network. The positions of neighbors in a perfect BCC crystal at the same density are shown in Fig. 6.7 (a). The second neighbors perfectly overlap with the split second peak of the EDIP $g(r)$, but the populations are wrong. The first two BCC peaks contains 8 and 6 neighbors, respectively, while the EDIP peaks contain roughly four each. So, if

[6]Once again, due to the small radius of the second peak, it is really a splitting of the first neighbors into two peaks, one covalent and one metallic.

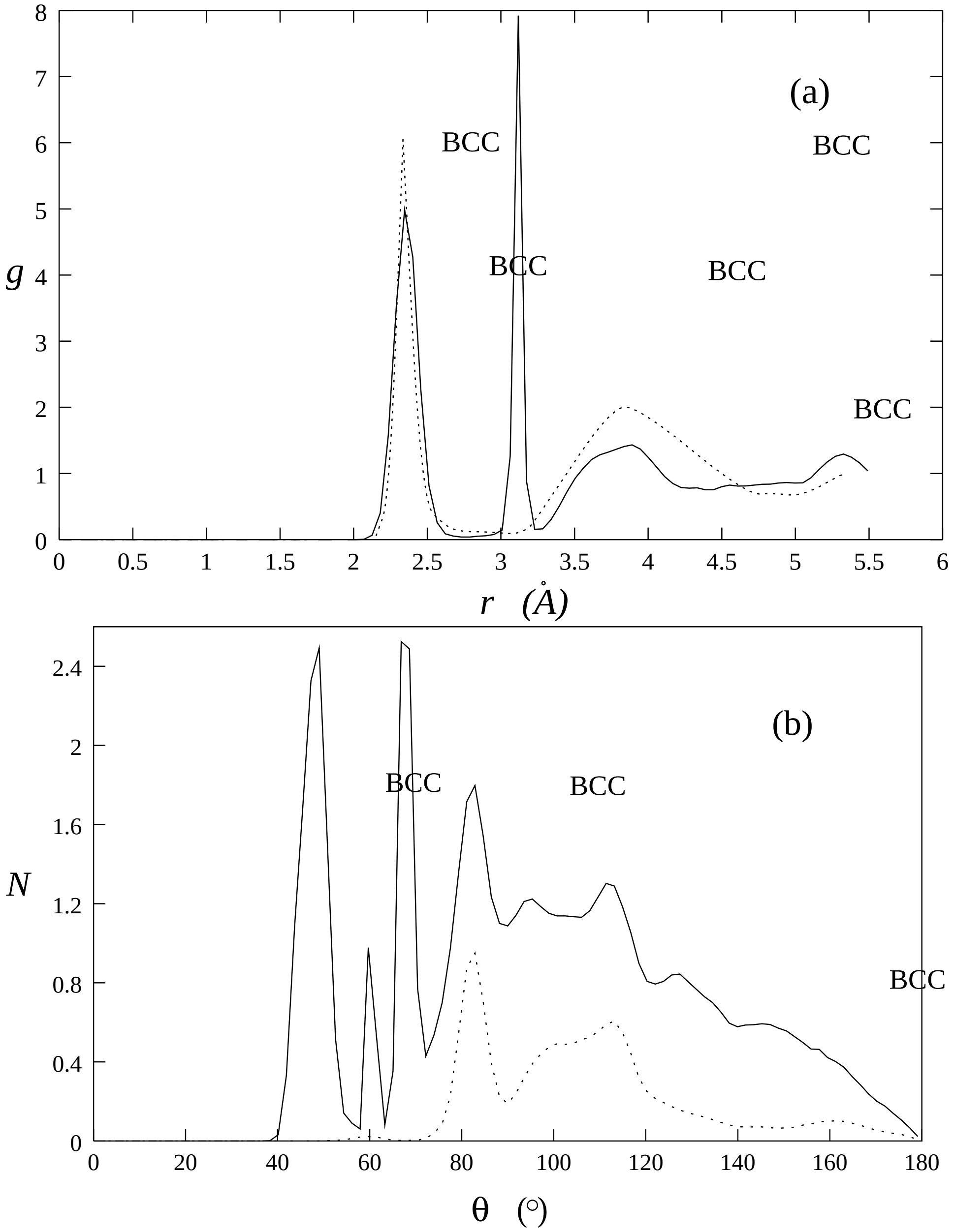

Figure 6.7: Structure of the EDIP amorphous phase a-EDIP-I obtained from quenching the liquid. In (a), the EDIP (solid line) and *ab initio* (dotted line) [153] pair correlation functions are compared, along with peaks in $g(r)$ for a BCC crystal at the same density. In (b), the EDIP bond angle distributions for $r < 2.83$ Å (dotted line) and $r < 3.26$ Å (solid line) are shown, and compared with BCC angles.

the EDIP amorphous has any BCC character, it must come from a splitting of the first peak into two four neighbor subshells.

In spite of the bizarre coordination shells, the pair correlation of the EDIP amorphous bears some resemblance to *ab initio* at larger distances. The second neighbor peak corresponding to the diamond second neighbor distance $r = 3.84$ Å is clearly seen, and the diamond third neighbor peak at $r = 4.5$ Å is absent. These physically correct features are also predicted by SW, but only EDIP predicts the diamond fourth neighbor peak at $r = 5.43$ Å in agreement with *ab initio* [134, 153]. However, note that the shapes and sizes of the peaks are incorrect, and EDIP predicts a flat plateau in $g(r)$ for $4.2 < r < 5.1$ Å.

In the liquid, the bond angles were quite well described in spite of the anomalous first peak splitting of $g(r)$, but bond angles reveal the EDIP quenched amorphous to have a fascinating, but completely unphysical structure. The *ab initio* bond angle distribution for a-Si is simply a bell curve of half-width 20° centered at the tetrahedral angle (with a tiny extra peak at 60°) [153]. The EDIP amorphous, on the other hand, has a remarkable degree of local orientational order, characteristic of a metallic crystal, as seen in Fig. 6.7 (b). As in the liquid case, the splitting of the first peak of $g(r)$ contains qualitatively different physics in the bond angle distributions as well. The bond angles for neighbors in the first peak of $g(r)$ with $r < 2.76$ Å, have an asymmetric peak at the tetrahedral angle, with a sharp decay on the large angle side, and a shoulder at 100° on the small angle side, indicating some sp^3 covalent bonding in the inner subshell, as in the liquid. There are also, however, a much larger and narrower peak at 83° and a small peak at at 160°, showing that the tetrahedra are systematically distorted to fit into a larger structural unit. The local coordinations in the first peak are 59% four-fold and 36% five-fold. The local structure of the five-fold coordinated inner subshell atoms may be similar to BCT5, which involves angles 86°, 106° and 148° (the first two are consistent with the data).

The full bond angle distribution, including the second subshell, is quite complicated. The distribution involves very sharp and tall peaks at small angles (48°, 60°, 68°, 83°) characteristic of metallic quasi-crystalline order. The sharp peak at 68° lies close to the BCC angle of 70.53° we mentioned earlier as an atom sitting on a tetrahedral face. The other strong small angle peak is close to the angle of 54.74° for an atom on a tetrahedral edge. The tetrahedral angle itself, also in the BCC crystal, is present in the outer subshell, and large angles also appear in the range $120 - 160°$, which are more or less absent in the first subshell. The structure does not have all the features of BCC, however, because angles near 180° are conspicuously missing. The many subsidiary peaks are probably related to averaging over a number of characteristic local structures. This is indicated by the broad distribution of local coordinations, which is peaked at 8 and 9 at 25% and 22%, respectively, with moderate numbers of atoms, around 15% each, at 7 and 10. The sharpness of the small angle peaks, however, suggests that these many different local structures contain similar building blocks. A typical structure, consistent with the data, might involve a BCC eight-atom coordination shell with four tetrahedral atoms pinched inward (first split peak of $g(r)$) and the other four moved outward (second split peak). In summary, although EDIP has not found the familiar tetrahedral structure of a-Si, it has, nevertheless, quenched into a curious random network with a different and more complicated local structure.

6.3.2 Another Amorphous Phase

The inability of the EDIP liquid to quench into the correct amorphous structure is not a major setback, since no other potential can do it either. Based on the extensive theoretical and fitting input involving sp^3-bonded structures, however, we would expect EDIP to be capable of describing a-Si if it could be coaxed into the correct structure. It seems reasonable that the complexity of the potential, which allows for strange phases like the diamond-BCC amorphous, would cause the free energy barrier between the quenched liquid and the highly-ordered (low entropy) tetrahedral network to be too large to overcome in a fast quench. A much less stringent test than quenching is to see what happens when a realistic amorphous sample is annealed with EDIP forces. Unfortunately, and perhaps surprisingly, the present version of EDIP fails this test: Not only is the correct structure dynamically unstable, but a spontaneous transition to another overcoordinated amorphous network occurs.

A realistic sample of a-Si can be generated using the bond-switching algorithm of Wooten, Winer and Weaire [155] or a molecular dynamics technique involving modified SW potentials (which generates the so-called "indirect amorphous") [133, 134]. In our tests we use a 216-atom sample created with the latter method (by N. Bernstein). The purity of the sample is improved by running at 300 K with double-strength three-body energy, and then annealing down to 100K with the unmodified SW potential. This structure is equilibrated with EDIP at 600 K and quenched down to 100 K in 18 ps, resulting in a phase we may call a-EDIP-II. The energy of a-EDIP-II, -4.45 eV/atom, is higher than a-EDIP-I by 0.05 eV/atom and higher than the diamond ground state by 0.21 eV/atom. The unphysical nature a-EDIP-II is suggested by the density *increase* of 5% versus the equilibrium crystal. As seen in Fig. 6.8, the pair correlation function retains some of the original structure, with a sharp isolated first peak at 2.38 Å, containing 4.17 neighbors, but as in the all disordered structures we have encountered, a sharp second peak (really a splitting of the first peak) appears at 3.12 Å, containing 1.71 more neighbors for a total coordination of 5.88. The distribution of local coordinations, given in Table 6.4, shows the predominance of five-fold and six-fold coordination with many larger coordinations also present, which accounts for the increased density. Of course, the coordination distributions from more realistic models are sharply peaked at 4. The a-EDIP-II bond angle distribution in Fig. 6.8 (b) is quite similar to that of a-EDIP-I, with all the same narrow, small-angle peaks, and the inner subshell structure resembling the liquid. The main difference between I and II is that the former has 4

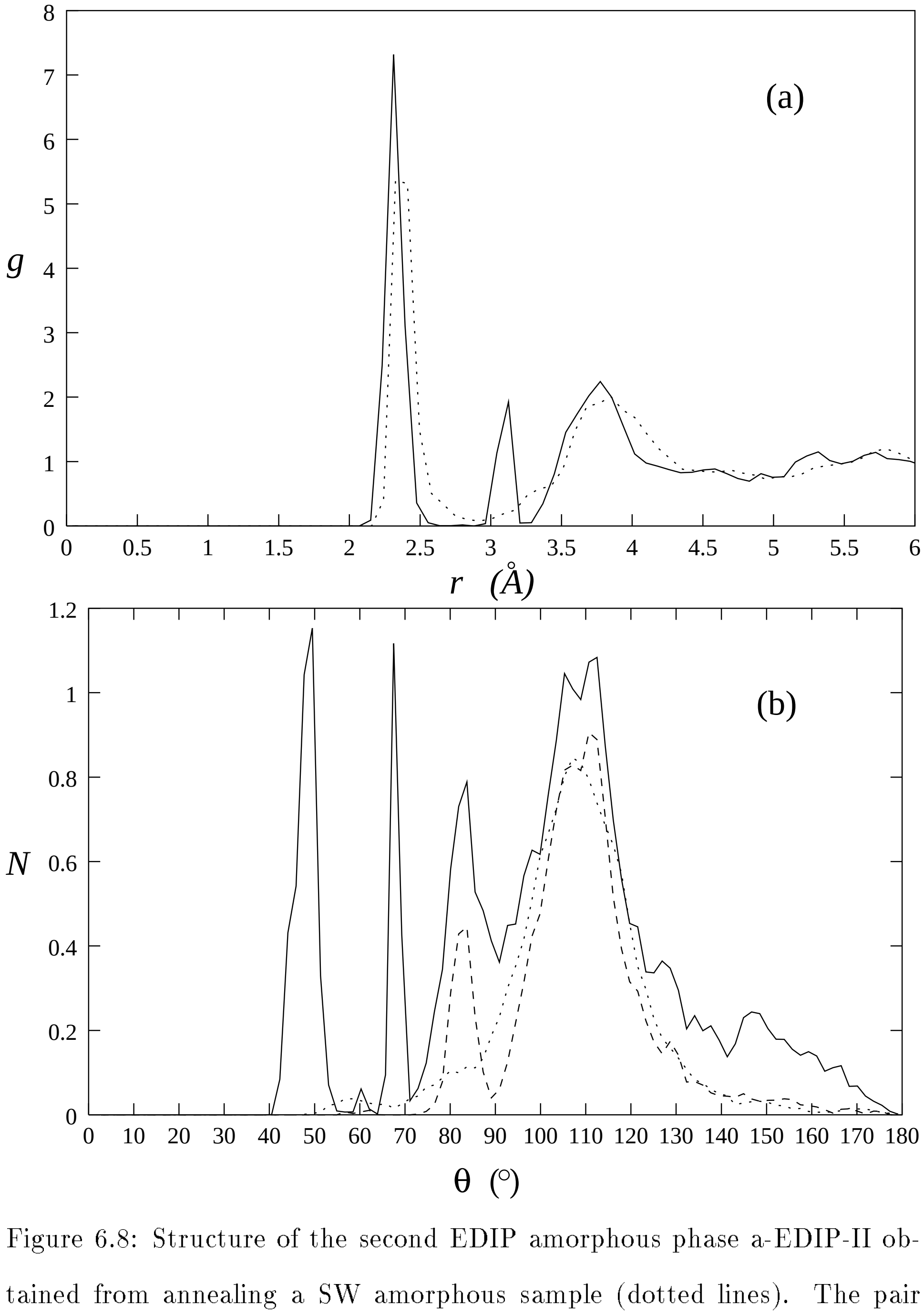

Figure 6.8: Structure of the second EDIP amorphous phase a-EDIP-II obtained from annealing a SW amorphous sample (dotted lines). The pair correlation function is shown in (a), and the bond angle distributions for the first (solid line) and first two (dashed line) peaks of $g(r)$.

Z	LDA	WWW	SW	EDIP
3	0.2	1.2	0.0	0.0
4	96.6	86.6	81.0	12.1
5	3.2	11.8	18.1	29.7
6	0.0	0.2	0.9	31.1
7	0.0	0.0	0.0	14.2
≥ 8	0.0	0.0	0.0	12.9

Table 6.4: Coordination statistics for amorphous structures generated by *ab initio* (LDA) dynamics with quenching [153], the Wooten-Winer-Weaire algorithm (WWW) [155], the indirect SW method (SW) [133], and annealing of the indirect SW structure with EDIP, a-EDIP-II.

instead of 2 neighbors on average in the outer, metallic subshell, and thus also has a much smaller volume ($V_I = 16.45$, $V_{II} = 19.00$ Å^3/atom). Otherwise various bonding arrangements are quite similar in the two amorphous phases.

As a last hope for EDIP to describe a-Si, we can try to build the amorphous phase by annealing a structure with fewer overcoordinated defects than the indirect SW sample. However, it turns out that a rather wide range of structures resembling a-Si transform into a-EDIP-II upon annealing, indicating that it is a free energy minimum with a broad basin of attraction. For example, a-EDIP-II emerges from annealing the low-density, four-fold coordinated structure generated by quenching the SW liquid with the SW potential modified for double-strength three-body energy (without the final step of annealing with the true SW potential). This starting point has greater volume and fewer overcoordinated atoms than the indirect SW amorphous, but it still relaxes into a-EDIP-II with EDIP forces, even at room temperature.

6.4 Thermal Stability of Bulk Defects

The results of the previous section lead us to question the range of transferability of the current version of EDIP because it sometimes transforms into unphysical structures in the presence of disorder, even at low temperatures. For pervasive long-range disorder, we have documented the collapse into a-EDIP-I from quenching the liquid and into a-EDIP-II from annealing a variety of low-temperature, amorphous states, characterized by tetrahedral short-range order. Consistent with the latter result, a-EDIP-II is formed when an amorphous-crystalline interface (created by N. Bernstein) is relaxed with EDIP at $T = 1000$ K, which should instead lead to crystallization in real silicon. These results are troublesome because they suggest that bulk crystalline defects, whose energies are exceptionally well described at $T = 0$, may be thermally unstable. Fortunately, that is not the case, as evidenced by a number of tests.

First of all, note that the splitting of the first peak of $g(r)$ present in all the unphysical disordered phases does not appear when heating a perfect diamond crystal all the way up to the melting point. Hence, the diamond structure is thermally stable against the kinds of unphysical, structural relaxations we have been examining. The same is true of some isolated point defects. For example, the unphysical rearrangements do not occur when a sample with a vacancy concentration of 0.05% is heated to the melting point. This concentration is many orders of magnitude larger than any observed in experiment. If the vacancy concentration is increased even further, then eventually there is sufficient disorder that regions of a-EDIP-II are nucleated. At 5% vacancy concentration, a-EDIP-II domains appear even at $T = 100$ K, but this is not problematic because such a high concentration implies non-isolated defects and voids, characteristic of a (very unphysical) foam. On the other hand, extensive tests of partial dislocation structures have not uncovered these kinds of problems, and in general, there appears to be no problem with finite temperature simulations of crystal defects right up to the melting point. Thus, we may claim EDIP to be a reliable potential for the simulation of bulk defects.

6.5 Prospects for Increased Transferability

Of course, we would like to be able to claim more. The theory behind the functional form should be capable of describing all the bulk phases, both covalent and metallic, ordered and disordered. Let us begin by addressing the amorphous, the most serious problem with the current version.

The only way to create and stabilize a more realistic amorphous (and avoid a-EDIP-I and a-EDIP-II) is apparently to modify the potential. We need to somehow eliminate the outer part of the split first peak of $g(r)$. Since this peak arises where the three-body force nearly vanishes but the pair attraction does not, it seems reasonable to try setting $b = a$ in the potential. This extends the range and increases the strength of three-body forces. Sure enough, when the indirect SW amorphous state is annealed with such a modified EDIP (without any refitting of the other parameters), a more reasonable structure results, as shown in Fig. 6.9 (dashed lines). It is a great improvement over the I and II phases described above, but is less realistic than the indirect SW amorphous. The isolated first peak contains 4.50 neighbors with covalent bond lengths, and the metallic second split peak is absent. An improvement over the indirect SW amorphous include the vanishing pair correlation in between the first two peaks, indicating less disorder in the random network, and the density, which is almost the same as the equilibrium diamond crystal. The bond angles are nicely peaked around tetrahedral, and the small angles peaks have mostly vanished. However, some new unphysical features appear, namely a broad peak at $60°$ and a small shoulder at $145°$. Nevertheless, with a simple change (suggested by the cohesive energy curves), namely increasing the three-body cutoff to the pair cutoff distance, we have salvaged reasonable behavior without any refitting of the potential. Of course, all other properties are altered, including elastic constants and defect formation energies, and there is no guarantee that the desirable

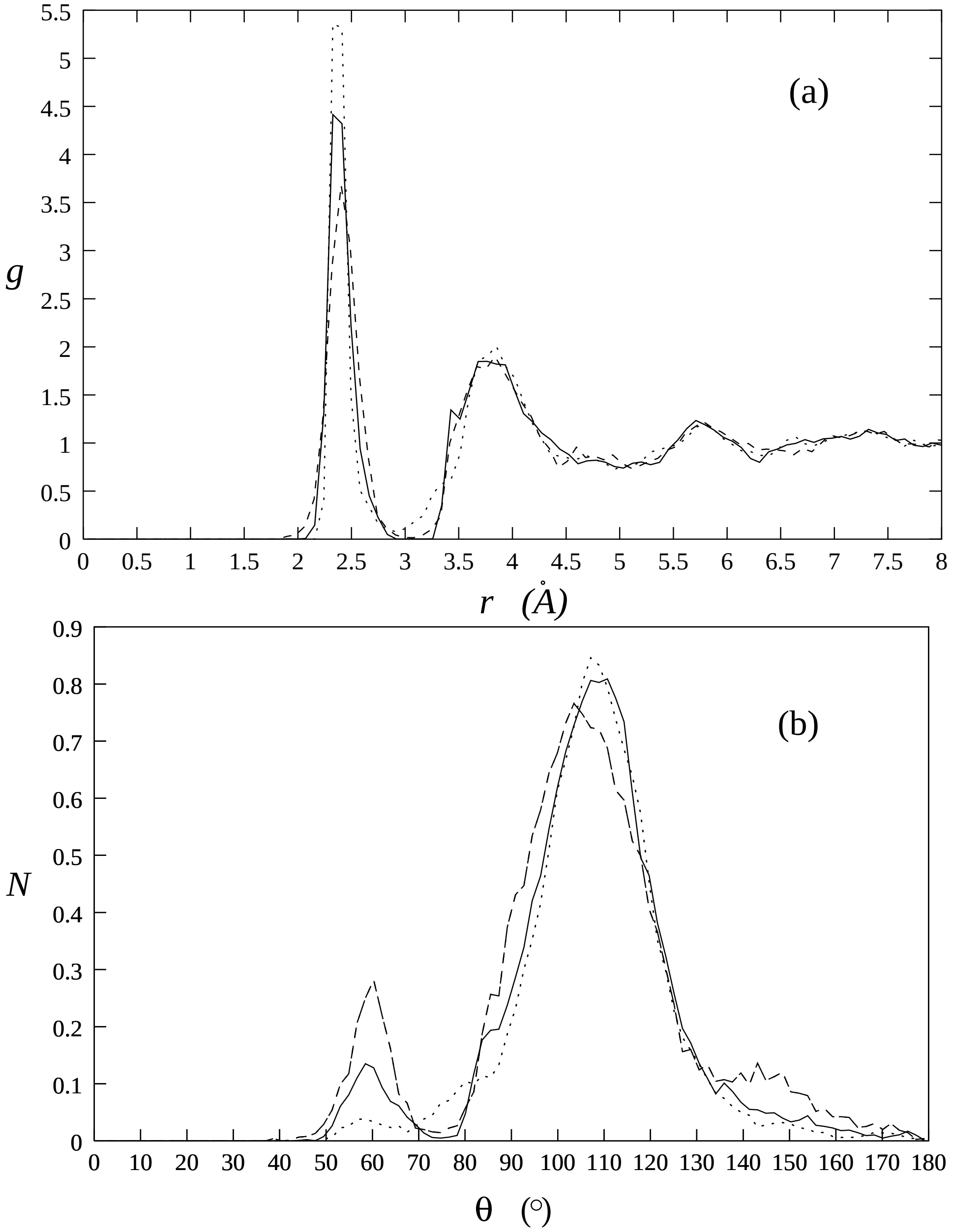

Figure 6.9: The pair correlation (a) and bond angle distribution (b) for amorphous phases of EDIP modified by using $H(x) = \lambda x^2$ with an extended three-body radial function (solid lines) and by simply setting $b = a$ (dashed lines), compared with the indirect SW structure (dotted lines).

features of the current version of EDIP can be preserved while fixing the problems with the amorphous phase. Indeed, recent experience in refitting the potential with the constraint of $b = a$ suggests that this goal is impossible with the current functional form.

To better understand the prospects for a simultaneous description of the amorphous phase and crystalline defects with EDIP, we explore modifications to the functional form that have minimal effect on bulk properties. The crucial gap in the theoretical foundation of EDIP explained in Section 5.1 is the shape of the angular function $H(x)$. The coordination dependence $x = (\cos\theta + \tau(Z))/w(Z)$ is supported by various arguments, but the particular choice $H_1(x) = \lambda(1 - \exp(-x^2))$, in spite of some appealing properties, has no real justification (aside from $H(0) = H'(0) = 0$ and $H''(0) > 0$).

Both a-EDIP-I and a-EDIP-II are characterized by the presence of small angles, $\theta < 90°$, which are mostly absent in the fitting database. If we could increase the penalty for small angles, then most of the database energies would not be affected. It turns out that (keeping all parameters the same) adding an additional penalty for small angles to the current version, $H(x) = H_1(x) + \eta x^m$ (for $m = 4, 6, 8$ and small η) does not solve the amorphous problems.

Drawing on the relative success of SW, it seems natural to try $H(x) = \lambda x^2$, which makes the EDIP form reduce to SW for $Z = 4$. This choice (with the original parameters of EDIP), keeps the same elastic constants and greatly increases the penalty for small angles. With this change, the amorphous improves dramatically (but not completely), indicating the great importance of the angular function. The small angle peaks of the bond angles reduce along with the split peak of the pair correlation, but none vanish.

A satisfactory amorphous can be produced with the EDIP functional form by combining the two helpful changes just described. First the angular function is changed to $H(x) = \lambda x^2$, and second the range of the three-body radial function $g(r)$ is extended. Unlike the exercise above where we simply set $b = a$, we now do the same with complementary changes to $g(r)$ to preserve values near $r = 2.35$ Å. Specifically, we choose a larger value $\gamma = 0.5$ (for reasons described in Section 5.2.3), and adjust the magnitude of $g(r)$ by multiplying the original λ by 1.225^2. This extends $g(r)$ to larger radii, $2.8 < r < 3.2$ Å, without affecting much the range, $2.2 < r < 2.6$ Å, present in most bulk defect structures. As shown Fig. 6.9 (solid lines), the pair correlation and bond angles are quite realistic with these modifications. The coordination under the first peak is 4.37, and 72% of atoms have four and 25% have five neighbors, which agrees with the *ab initio* data in Table 6.4 almost as well SW.

Although it has not been fitted to anything explicitly (other than the elastic constants and diamond structure, which we have preserved), it is interesting to test our amorphous-modified EDIP for the liquid. It does not perform well, but behaves differently than the original version. The structure at $T = 2500$ K and $P = 0$ shown in Fig. 6.10 is reminiscent of a hot amorphous phase. The pair correlation has a fairly well-isolated first peak containing 5.67 neighbors, and the bond angle distribution is bimodal with peaks at 60° and 90°. With this example we see the difficulty of simultaneously describing all the bulk phases, consistent with the observations of Ding and Andersen from their work in applying the SW potential to germanium [136].

The difficulty of the problem is also suggested by our analysis of the original EDIP liquid and amorphous phases. In the former, we have determined that the split first peak of $g(r)$ should be merged, preserving the mixture of covalent and metallic bonds, particularly the local structure of Fig. 6.5. In the latter, we must somehow achieve the opposite. In the amorphous the outer shell of first peak must be removed, leaving only covalent, tetrahedral atoms. Thus, in order for EDIP to describe both the liquid and amorphous phases, a delicate balance must be struck between the tendencies for metallic and covalent bonding.

Nevertheless, EDIP shows exceptional promise as general and transferable model for bulk phases and defects. We have seen that its functional form is much more flexible

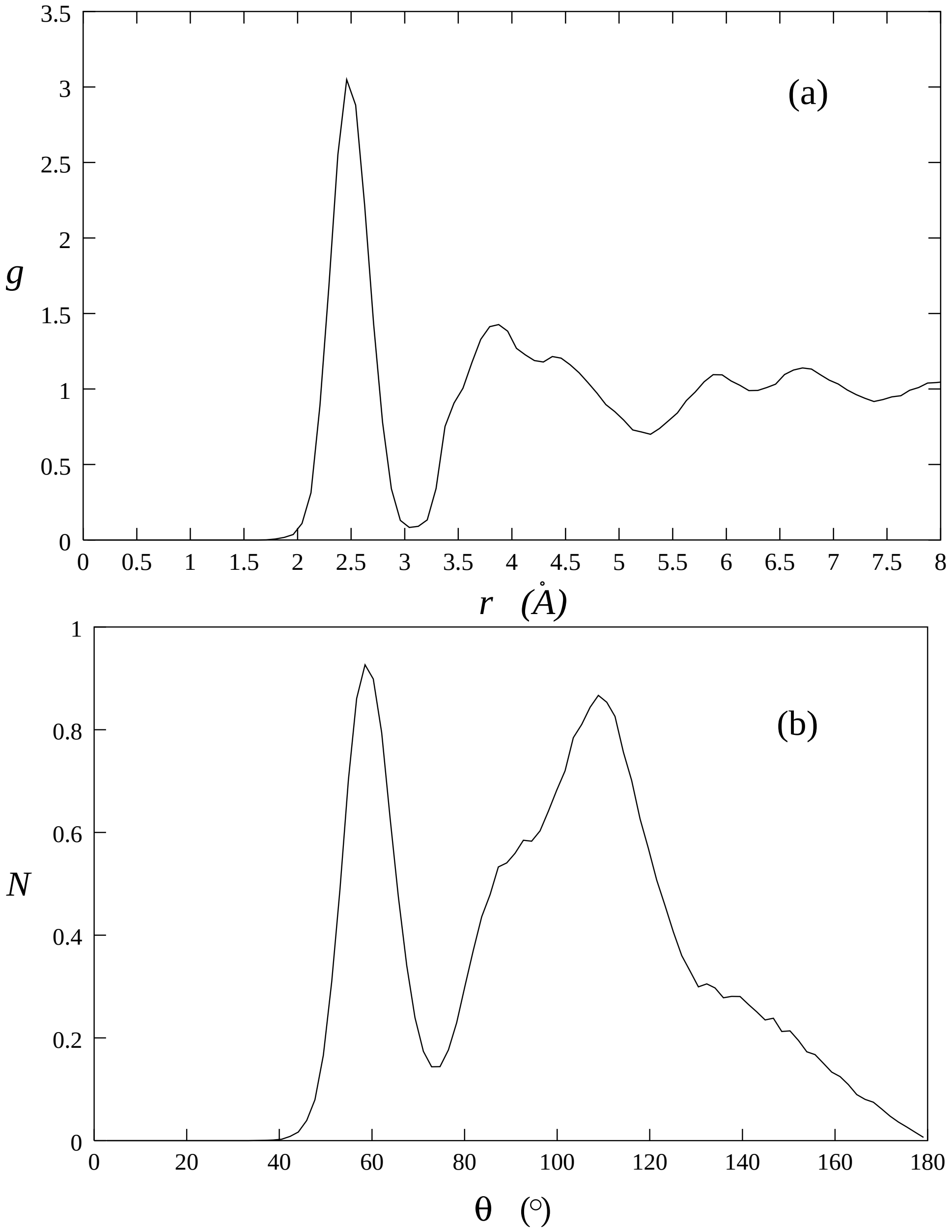

Figure 6.10: The liquid pair correlation (a) and bond angle distribution (b) for EDIP modified by using $H(x) = \lambda x^2$ with an extended three-body radial function.

than others, in spite of its comparable computational efficiency and small number of parameters. The examples in this section provide some guidance for refitting the potential for disordered phases, and it is likely that the amorphous and liquid can both be accommodated with minor changes to the functional form and some refitting.

Chapter 7

Conclusion

> Of course, every new potential is claimed by its originators to be superior, i.e., more accurate and/or more transferable than its predecessors. While these claims are often valid to some extent, such improvements are almost always achieved by sacrificing other properties. Also, very often it is not truly clear what causes the better description. Is it due simply to a more flexible functional form and/or fitting strategy or does the new potential really give a better description of covalent bonding?
>
> – H. Balamane, T. Halicioglu and W. A. Tiller [19]

Let us objectively discuss our successes and failures and then look forward to the future of EDIP and empirical potentials for covalent solids, in general. Throughout this thesis we have answered many of our motivating questions affirmatively. By reviving and improving several analytic techniques from the literature of solid state physics, we have established various facts concerning the functional form of interatomic forces in the prototypical case of Si directly from *ab initio* calculations, which have proven useful in designing a transferable fitted potential for silicon bulk phases and crystal defects.

Through elastic constant analysis we have studied forces mediated by sp^2 and sp^3 hybrid bonds in covalent structures. In the case of the Harrison model for diamond elasticity, we have demonstrated that a simple, underdetermined functional form can fit a nontrivial manifold on the Born-Oppenheimer energy surface almost perfectly, as evidenced by the elastic constant relation, $4C_{11} + 5C_{12} = 9C_{44}^o$, which is satisfied by experimental and *ab initio* data for Si. We interpret this success as validation of the Rigid Hybrid Approximation for *any* elastic deformation without internal relaxation. For shear strains with relaxation, measured by a nonzero Chelikowsky dangling bond vector, the Harrison model fails because it does not describe rehybridization. We have also confirmed that second neighbor forces in the diamond lattice are very weak, and in the case of a three-constant model have demonstrated that adding a degree of freedom does not guarantee a better fit if the functional form is wrong. Interesting comparisons between different hybrid covalent bonds have also been made by analyzing a Harrison-like model for the elastic constants of a hexagonal plane. Our *ab initio* calculations for Si reveal that sp^2 hybrids have a greater radial force constant but a weaker angular force constant than sp^3 hybrids, an important and counterintuitive result.

In order to explore global trends in bonding across bulk structures, we have performed the first meaningful inversions of cohesive energy curves for a covalent solid. This is accomplished by understanding and solving problems with long-range forces and deriving formulae for many-body interactions. In response to one of our motivating questions, it is indeed possible to derive competitive many-body potentials directly from *ab initio* data without any adjustable parameters. By looking at different bulk phases, we have also exposed environment dependence, showing that the bond order form of the pair interaction is in excellent agreement with theory.

Aside from gaining physical insight through inversion, we have also developed some new mathematics. With our many-body formulae in Chapter 4 and Appendix B, several classes of *nonlinear* inverse problems are solved. The central idea of recursion also finds interesting applications in number theory related to the Möbius Inversion Formula, as described in Appendix C.

On a more practical note, building upon this work we have proposed a functional form for interatomic forces in covalent solids with only 13 fitting parameters, called the Environment-Dependent Interatomic Potential. It blends the desirable features of the Tersoff and SW models we have identified theoretically and includes a new environment-dependent angular function, which adapts the angular stiffness and favored angle to model rehybridization and metalization. An important point is that force evaluation with EDIP is as fast as with much simpler models. A fitted EDIP for bulk defects in Si is remarkably realistic for diamond elasticity and a wide range of defect structures not in the fitting database, including generalized stacking faults and reconstructed partial dislocation cores. The liquid is rather well described, aside from the unphysical splitting of the first neighbor peak of $g(r)$, but the amorphous phase is not correctly modeled by the current version. However, we have identified the sources of these problems and have shown how they can be corrected. With some additional work, it is likely that an EDIP for silicon will provide a superior description of the important bulk phases and defects.

Although we are surely guilty of overstating our successes to some degree, we have made a sincere effort to address Balamane's criticisms quoted above. Our testing has been much more extensive than any other potential prior to publication. In fact, had we stopped testing before looking at disordered phases, we might have thought we had stumbled upon *the* potential for bulk Si from the defect results. It is our goal to thoroughly understand the behavior of our potential, so future researchers can use it with confidence, safely warned about its limitations.

We have also taken unprecedented measures to help us interpret our successes and avoid the ambiguity of blind fitting schemes. In spite of its sophistication, EDIP has hardly any more degrees of freedom than the simplest models, so its successes cannot be due to increased flexibility. Our fitting strategy has also been kept fairly simple, with clear focus on a particular class of environment (bulk crystal defects) that is within our theoretically predicted range of validity. In contrast, other potentials are repeatedly extended to situations where there is no reason to expect success by simply adjusting arbitrary fitting parameters. As far as interpretation in terms of chemical bonding goes, we have demonstrated agreement with inversions of *ab initio* data in several ways, and in those cases success is not merely a matter of luck in fitting. In the majority of cases, however, we must admit that, aside from fitting and testing, we cannot carefully validate many aspects of the EDIP functional form, which surely are inadequate for complete transferability. Nevertheless, we are holding our work to high standards of theoretical and practical validation, because our overall aim in working with the one of the most difficult and extensively studied materials is to understand the general limitations of empirical interatomic potentials.

So, how will we know when to stop working on Si? That is a difficult question, but the answer must surely be, not yet. At this point, many researchers have given up on improving the description of Si, and have moved on to other covalent materials (like Ge, C, S, F, SiF, SiO_2, GeSe, SiN, ...) where less work has been done and the standards of accuracy are much lower. It is certainly important to study these materials, but given the difficulty in describing Si under close scrutiny, it is hard to believe that poorly tested potentials for less well understood materials can be trusted enough to generate realistic simulations. Still, questionable physical validity has not stopped the growing tide of large-scale atomistic simulations, fueled by growing excitement over advances in high performance computation.

In going to new materials that are less well understood, the methods developed in this thesis should be quite useful. For example, the environment dependence of the bond order could be checked for related covalent elemental solids and alloys, and our elastic constant relations could be used to compare angular forces and bond strengths for different hybrid bonds. These properties should be qualitatively similar to Si, but other materials will have important quantitative differences leading to different structural preferences. The EDIP functional form, since it contains environment dependence for metallic bonding and different covalent hybridizations, may provide a unified way to describe all covalent materials. It may be possible to use inversion and elastic constant results to simply rescale the parameters of the Si version of EDIP for other materials to obtain reasonable potentials. If this works, we can claim we have truly learned some general features of bonding in covalent solids.

Even if the dream of quantitatively accurate atomistic simulations with empirical potentials is never realized, there will always be a crucial role for potentials to play in materials science. Compared with accurate quantum-mechanical treatments, empirical potentials provide a means to explore the qualitative effect of going to larger system sizes or longer times. This capability is necessary, for example, to evaluate entropic contributions to free energies, so that predictions of *ab initio* energy calculations of atomistic mechanisms can be extended to finite temperatures. Another important use of empirical potentials is to quickly probe phase space looking for a small set of candidate atomic mechanisms to be studied quantitatively with *ab initio* methods. In other cases, where quantitative comparison with experiment is not needed, the essential physics of a process may be contained in simple models (*e.g.* phase transitions of the hard sphere model in statistical mechanics), so for certain general theories of materials phenomena, empirical potentials may be sufficient for qualitative understanding.

Beyond these practical uses, empirical potentials still have the power to dictate our conceptual understanding of chemical bonding. The concepts of pair bonds and angular forces developed through the models of Born, Harrison and Stillinger-Weber define the way we think about covalent materials. The language by which we understand the results of *ab initio* calculations is influenced by these artificial but useful theoretical constructs: a tendency to have as many nearest neighbors as possible at a preferred distance (and thus maximize the number of unstrained bonds) is countered by an aversion to inappropriate angles. The balance between these competing effects helps us understand atomic relaxations and motion. The Tersoff family of potentials introduces the next crucial concept, that the strength and length of a bond depends on its environment, weakening and lengthening as coordination is increased. The main conceptual contribution of EDIP is the idea that angular forces also depend on the environment, weakening with increasing coordination and shifting the preferred angle depending on the number of neighbors. Environment dependence is the key to understand bonding preferences in defect structures and disordered phases, where different coordinations can arise. It also gives a unified view of competing covalent phases, like diamond and graphite in the case of carbon. The next step will be to understand what kind of environment dependence is needed for surfaces and small clusters, which is beyond the scope of this thesis. An important part of this task would be an analysis of π-bonding, which we have safely ignored in this work. At least in the prerequisite case of bulk material, we have contributed to the theoretical understanding of interatomic forces in covalent solids.

Bibliography

[1] R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I (Addison-Wesley, Reading, Massachusetts, 1963) p. 1-2.

[2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964); W. Kohn and L. J. Sham, Phys. Rev. **140**, 1133 (1965).

[3] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[4] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[5] A. Cottrell, Von Hippel Award address, Materials Research Society Bulletin **22**, No. 5, 15 (1997).

[6] A. Messiah, *Quantum Mechanics*, Vol. II (John Wiley, New York, 1976), p. 781.

[7] R. Car and M. Parinello, Phys. Rev. Lett. **55**, 2471 (1985).

[8] A. Voter, Materials Research Society Bulletin *21*, No. 2, 17 (1996).

[9] M. Born, Ann. Physik **44**, 605 (1914).

[10] J. de Launay, Solid State Physics **2**, 220 (1956).

[11] H. B. Huntington, Solid State Physics **7**, 213 (1958).

[12] W. A. Harrison, Ph.D. Dissertation, University of Illinois, Urbana, Illinois (1956).

[13] F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).

[14] S. Foiles, Materials Research Society Bulletin **21**, No. 2, 24 (1996).

[15] M. S. Tang, C. Z. Wang, C. T. Chan and K. M. Ho, Phys. Rev. B **53**, 979 (1996); C. Z. Wang, K. M. Ho and C. T. Chan, Phys. Rev. Lett. **70**, 611 (1993).

[16] G. C. Abell, Phys. Rev. B **31**, 6184 (1985).

[17] A. E. Carlsson, in *Solid State Physics: Advances in Research and Applications*, edited by H. Ehrenreich and D. Turnbull (Academic, New York, 1990), **43**, p. 1.

[18] J. Tersoff, Phys. Rev. Lett. **56**, 632 (1986).

[19] H. Balamane, T. Halicioglu, and W. A. Tiller, Phys. Rev. B **46**, 2250 (1992).

[20] D. G. Pettifor, Springer Proc. in Physics **48**, 64 (1990).

[21] P. Alinaghian, S. R. Nishitani and D. G. Pettifor, Phil. Mag. B **69**, 889 (1994);

[22] J. R. Chelikowsky J. C. Phillips, M. Kamal and M. Strauss, Phys. Rev. Lett. **62**, 292 (1989); J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B **41** 5735 (1990); J. R. Chelikowsky, K. M. Glassford and J. C. Phillips, **44** 1538 (1991).

[23] A. Guinier, *The Structure of Matter* (Edward Arnold, New York, 1984).

[24] T. Halicioglu, H. O. Pamuk and S. Erkoc, Phys. Stat. Sol. B **149**, 81 (1988).

[25] Z. Q. Wang and D. Stroud, Phys. Rev. B **38**, 1384 (1988).

[26] J. Tersoff, Phys. Rev. Lett. **61**, 2879 (1988).

[27] E. P. Andribet, J. Dominguez-Vasquez, A. M. C. Perez-Martin, E. V. Alonso and J. J. Jiminez-Rodriguez, Nucl. Instrum. Methods Phys. Res. B **115**, 501 (1995).

[28] F. H. Stillinger and T. A. Weber, J. Chem. Phys. **88**, 5123 (1988).

[29] F. H. Stillinger, T. A. Weber and R. A. Laviolette, J. Chem. Phys. **85**, 6460 (1986); F. H. Stillinger and T. A. Weber, **91**, 4899 (1987).

[30] T. Ito, K. E. Khor and S. Das Sarma, Phys. Rev. B **40**, 9715 (1989); K. E. Khor and S. Das Sarma, J. Vac. Sci. Technol. B **10**, 1994 (1992).

[31] J. Tersoff, Phys. Rev. B **39**, 5566 (1989).

[32] P. C. Kelires and J. Tersoff, Phys. Rev. Lett. **63**, 1164 (1989).

[33] F. H. Stillinger and T. A. Weber, Phys. Rev. Lett. **62**, 2144 (1989).

[34] P. Vashishta, L. Rajiv, K. Kalia, J. P. Rino and I. Ebbsjo, Phys. Rev. B **41**, 12197 (1990).

[35] P. Vashishta, R. Kalia, G. A. Antonio and I. Ebbsjo, Phys. Rev. Lett. **62**, 1651 (1989).

[36] E. Kaxiras, Comp. Mater. Sci. **6**, 158 (1996).

[37] J. Tersoff, Phys. Rev. B **37**, 6991 (1988).

[38] J. Tersoff, Phys. Rev. B **38**, 9902 (1988).

[39] J. Ferrante, J. R. Smith, and J. H. Rose, Phys. Rev. Lett. **50**, 1385 (1983); J. H. Rose, J. R. Smith and J. Ferrante, Phys. Rev. B **28**, 1835 (1983).

[40] M. Ishimaru, K. Yoshida, and T. Motooka, Phys. Rev. B **53**, 7176 (1996).

[41] R. Biswas and D. R. Hamann, Phys. Rev. Lett. **55**, 2001 (1985); Phys. Rev. B **36**, 6434 (1987).

[42] E. Kaxiras and K. Pandey, Phys. Rev. B **38**, 12736 (1988).

[43] J. R. Chelikowsky, Phys. Rev. Lett. **60**, 2669 (1988).

[44] B. W. Dodson, Phys. Rev. B **35**, 2795 (1987).

[45] K. E. Khor and S. Das Sarma, Phys. Rev. B **38**, 3318 (1988).

[46] K. E. Khor and S. Das Sarma, Phys. Rev. B **39**, 1188 (1989); **40**, 1319 (1989).

[47] D. W. Brenner, Phys. Rev. Lett. **63**, 1022 (1989).

[48] M. I. Baskes, Phys. Rev. Lett. **59**, 2666 (1987); M. I. Baskes, J. S. Nelson and A. F. Wright, Phys. Rev. B **40**, 6085 (1989).

[49] J. Wang and A. Rockett, Phys. Rev. B **43**, 12571 (1991).

[50] P. N. Keating, Phys. Rev. **145**, 637 (1966).

[51] D. W. Brenner and B. J. Garrison, Phys. Rev. B **34**, 1304 (1986).

[52] A. M. Stoneham, V. T. B. Torres, P. M. Masri and H. R. Schober, Phil. Mag. A **58**, 93 (1988).

[53] A. R. Al-Derzi, R. L. Johnston, J. N. Murrel and J. A. Rodriguez-Ruiz, Mol. Phys. **73**, 265 (1991); J. N. Murrel and J. A. Rodriguez-Ruiz, **71**, 823 (1990).

[54] E. M. Pearson, T. Takai, T. Halicioglu and W. A. Tiller, J. Crystal Growth **70**, 33 (1984).

[55] A. D. Mistriotis, N. Flytzanis, and S. C. Farantos, Phys. Rev. B **39**, 1212 (1989).

[56] B. Bolding and H. Anderson, Phys. Rev. B **41**, 10568 (1990).

[57] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).

[58] D. F. Richards and J. B. Adams, in *Grand Challenges in Computer Simulation*, Proceedings High Performance Computing '95, ed. by A. Tentner, 218 (1995).

[59] D. F. Richards, J. B. Adams, J. Zhu, L. Yang, and C. Mailhiot, Bull. Am. Phys. Soc. **41**, 264 (1996).

[60] J. S. McCarly and S. T. Pantelides, Bull. Am. Phys. Soc. **41**, 264 (1996).

[61] J. Harris, Phys. Rev. B **31**, 1770 (1985); A. P. Sutton, M. W. Finnis, D. G. Pettifor and Y. Ohta, J. Phys. C **21**, 35 (1988).

[62] A. P. Sutton, M. W. Finnis, D. G. Pettifor and Y. Ohta, J. Phys. C **21**, 35 (1988).

[63] D. Pettifor, *Bonding and Structure of Molecules and Solids* (Clarendon Press, Oxford, 1995).

[64] A. E. Carlsson, P. A. Fedders and C. W. Myles, Phys. Rev. B **41**, 1247 (1990).

[65] A. E. Carlsson, Springer Proc. in Phys. **48**, 257 (1990).

[66] A. E. Carlsson, Phys. Rev. B **32**, 4866 (1985); A. E. Carlsson and N. W. Ashcroft, **27**, 2101 (1983).

[67] A. P. Horsfield, A. M. Bratkovsky, M. Fearn, D. G. Pettifor and M. Aoki, Phys. Rev. B **53**, 12694 (1996).

[68] W. A. Harrison, Phys. Rev. B, **41**, 6008 (1990).

[69] H. Metiu, Mini-Workshop on Interatomic Potentials and Multi-Scale Modeling, Institute for Theoretical Physics, University of Santa Barbara, June 5-7 (1997).

[70] M. Born, in *Lattice Dynamics*, ed. by R. F. Wallis (Pergamon, New York, 1965), p.1.

[71] E. R. Cowley, Phys. Rev. Lett. **60**, 2379 (1988).

[72] S. R. Nishitani, P. Alinaghian, C. Hausleitner and D. G. Pettifor, Phil. Mag. Lett. **69**, 177 (1994).

[73] A. E. H. Love, *Mathematical Theory of Elasticity*, 4th ed. (Cambridge University Press, 1927).

[74] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices* (Clarendon Press, Oxford, 1954).

[75] A. H. Cottrell, *The Mechanical Properties of Matter* (Kreiger, New York, 1981).

[76] W. A. Harrison, *Electronic Structure and the Properties of Solids* (Dover, N.Y. 1980).

[77] M. J. P. Musgrave and J. A. Pople, Proc. Roy. Soc. A **268**, 474 (1962).

[78] G. Simmons and H. Wang, *Single Crystal Elastic Constants and Calculated Aggregate Properties: A Handbook* (MIT, Cambridge, MA, 1971).

[79] O. N. Nielsen and R. M. Martin, Phys. Rev. B **32**, 3792 (1985).

[80] N. Bernstein and E. Kaxiras, submitted to Phys. Rev. B; in *Materials Theory, Simulations and Parallel Algorithms*, ed. by E. Kaxiras, J. Joannopoulos, P. Vashista, and R. Kalia, Materials Research Society Symposia Proceedings, **408** (M. R. S., Pittsburgh, 1996), 55.

[81] L. Kleinman, Phys. Rev. **128**, 2614 (1962).

[82] C. S. G. Cousins, L. Gerward, J. Staun Olsen, B. Selsmark and B. J. Sheldon, J. Appl. Crystallogr. **15**, 154 (1982).

[83] A. Segmuller, Phys. Kondens. Mater. **3**, 18 (1964).

[84] A. Segmuller and H. R. Neyer, Phys. Kondens. Mater., Phys. Kondens. Mater. **4**, 63 (1965).

[85] J. F. Nye, *Physical Properties of Crystals* (Clarendon Press, Oxford, 1957).

[86] W. A. Harrison and J. C. Phillips, Phys. Rev. Lett. **33**, 410 (1974).

[87] D. J. Chadi, Phys. Rev. B **29**, 785 (1984).

[88] A. P. P. Nicholson and D. J. Bacon, J. Phys. C: Solid State Phys. **10**, 2295 (1977).

[89] H. Poincare, *The Value of Science* (Dover, New York, 1958), p.76.

[90] A. Carlsson, C. Gelatt, and H. Ehrenreich, Phil. Mag. A **41** (1980).

[91] B. Friedman, *Principles and Techniques of Applied Mathematics* (Wiley, New York, 1956).

[92] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College Publishers, New York, 1976).

[93] N.-X. Chen, Phys. Rev. Lett. **64**, 1193 (1990).

[94] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers* (Oxford Univ. Press, Oxford, 1979), 5th ed.

[95] R. A. Dean, *Elements of Abstract Algebra* (John Wiley, New York, 1966).

[96] J. Maddox, Nature **344**, 377 (1990).

[97] N.-X. Chen and G.-B. Ren, Phys. Rev. B **45**, 8177 (1992).

[98] J. Wang, K. Zhang, and X. Xie, J. Phys. C **6**, 989 (1994).

[99] N.-X. Chen, Z.-D. Chen and Y.-C. Wei, Phys. Rev. E **55**, R5 (1997).

[100] Q. Xie and N.-X. Chen, Phys. Rev. B **51**, 15856 (1995).

[101] M. W. Finnis and J. E. Sinclair, Phil. Mag. A **50**, 45 (1984).

[102] N. Moll, M. Bockstedte, M. Fuchs, E. Pehlke, and M. Scheffler, Phys. Rev. B **52**, 2550 (1995).

[103] M. T. Yin, Phys. Rev. B **30**, 1773 (1984).

[104] M. Yin and M. Cohen, Phys. Rev. B **26**, 5668 (1982).

[105] M. Mehl and L. Boyer, Phys. Rev. B **43**, 9498 (1991);

[106] E. Kaxiras and L. Boyer, Phys. Rev. B **50**, 1535 (1994).

[107] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory* (Pergamon, New York, 1977).

[108] L. Pauling, *The Nature of the Chemical Bond* (Cornell University, Ithaca, 1960).

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

[110] J. R. Chelikowsky, J. C. Phillips, M. Kamal, and M. Strauss, Phys. Rev. Lett. **62**, 292 (1989).

[111] K. C. Pandey, Phys. Rev. Lett. **57**, 2287 (1986).

[112] M. T. Yin and M. L. Cohen, Phys. Rev. B **24**, 2303 (1981).

[113] I. P. Batra, Phys. Rev. B **41**, 5048 (1990).

[114] M. Z. Bazant, E. Kaxiras and J. F. Justo, submitted to Phys. Rev. B (1997).

[115] J. F. Justo, M. Z. Bazant, E. Kaxiras, V. V. Bulatov and S. Yip, Materials Research Society Proceedings, Spring Meeting Symposium E (Materials Research Society, Pittsburgh, 1997).

[116] S. Ismail-Beigi and E. Kaxiras, private communication (1993).

[117] D. W. Brenner, M.R.S. Bulletin **21** no. 2, 36 (1996).

[118] E. Kaxiras and L. L. Boyer, Modeling Simul. Mater. Sci. Eng. **1**, 91 (1992).

[119] K. P. Huber and G. Herzberg, *Constants of Diatomic Molecules* (Van Norstrand, New York, 1979).

[120] M. S. Duesbery, D. J. Michel, E. Kaxiras and B. Joos, in *Defects in Materials*, Materials Research Society Proceedings **109**, ed. by P. D. Bristowe, J. E. Epperson, J. E. Griffith and Z. Lillenthal-Weber (Materials Research Society, Pittsburgh, 1991), p. 125.

[121] E. Kaxiras and M. S. Duesbery, Phys. Rev. Lett. **70**, 3752 (1993).

[122] Y. Bar-Yam and J. D. Joannopolous, Phys. Rev. Lett. **52**, 1129 (1984).

[123] P.J. Kelly and R. Car, Phys. Rev. B **45**, 6543 (1992).

[124] H. Seong and L. J. Lewis, Phys. Rev. B **53**, 9791 (1996).

[125] I. Kwon, R. Biswas, C. Z. Wang, K. M. Ho and C. M. Soukoulis, Phys. Rev. B **49**, 7242 (1994).

[126] J. Justo, *Atomistics of Dislocation Mobility in Silicon: Core Reconstruction and Mechanisms* (Ph.D Thesis, Department of Nuclear Engineering, Massachussetts Institute of Technology, 1997).

[127] P. E. Blöchl, E. Smargiassi, R. Car, D. B. Laks, W. Andreoni, and S. T. Pantelides, Phys. Rev. Lett. **70**, 2435 (1993).

[128] J. P. Hirthe and J. Lothe, *Theory of Dislocations* (Wiley, New York, 1982).

[129] J. R. K. Bigger *et. al.*, Phys. Rev. Lett. **69**, 2224 (1992).

[130] M. S. Duesbery, B. Joos, and D. J. Michel, Phys. Rev. B **43**, 5143 (1991).

[131] T. A. Arias, V. V. Bulatov, S. Yip, and A. S. Argon (to be published).

[132] D. L. Goodstein, *States of Matter* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[133] J. Q. Broughton and X. P. Li, Phys. Rev. B **35**, 9120 (1987).

[134] W. D. Luedtke and U. Landman, Phys. Rev. B **37**, 4656 (1988).

[135] M. D. Kluge, J. R. Ray and A. Rahman, Phys. Rev. B **36**, 4234 (1987).

[136] K. Ding and H. C. Andersen, Phys. Rev. B **34**, 6987 (1986).

[137] J. M. Holender and G. J. Morgan, J. Phys.: Condens. Matter **3**, 1947 (1991).

[138] P. C. Kelires and J. Tersoff, Phys. Rev. Lett. **61**, 562 (1988).

[139] P. Tamayo, J. P. Mesirov and B. M. Boghosian, Proceedings Supercomputing '91 (IEEE Comput. Soc. Press, Los Alamitos, CA, 1991), p. 462.

[140] A. Melcuk, R. C. Giles and H. Gould, Computers in Physics May/Jun, 311 (1991).

[141] R. C. Giles and P. Tamayo, Proceedings. Scalable High Performance Computing Conference SHPCC-92 (IEEE Comput. Soc. Press, Los Alamitos, CA, 1992), p. 240.

[142] P. S. Lohmdahl, P. Tamayo, N. Gronbech-Jensen and D. M. Beazley, Proceedings of Supercomputing '93 (IEEE Comput. Soc. Press, Los Alamitos, CA, 1993), p. 520.

[143] D. M. Beazley, P. S. Lohmdahl, N. Gronbech-Jensen and P. Tamayo, Proceedings Eighth International Parallel Processing Symposium (IEEE Comput. Soc. Press, Los Alamitos, CA, 1994), p. 800.

[144] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).

[145] H. C. Andersen, J. Chem Phys. **72**, 2384 (1980).

[146] M. Parinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).

[147] G. Bennetin, L. Galgani, and J.-M. Strelcyn, Phys. Rev. A, **14**, 2338 (1976); G. Bennetin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, Meccanica, **15**, 9 (1980).

[148] J. Tobochnik and H. Gould, Computers in Physics, Nov/Dec, 86 (1989).

[149] V. M. Glazov, S. N. Chizhevskaya and N. N. Glagoleva, *Liquid Semiconductors* (Plenum, New York, 1969), Ch. 3.

[150] M. Davidovic, M. Stojic and D. Jovic, J. Phys. C: Solid State Phys. **16**, 2053 (1983).

[151] I. Stich, R. Car and M. Parinello, Phys. Rev. Lett. **63**, 2240 (1989); Phys. Rev. B **44**, 4262 (1991).

[152] R. Biswas, G. S. Grest and C. M. Soukoulis, Phys. Rev. B **36**, 7437 (1987).

[153] I. Stich, R. Car and M. Parinello, Phys. Rev. B **44**, 11,092 (1991).

[154] N. W. Ashcroft, Il Nuovo Cimento D **12**, 597 (1990).

[155] F. Wooten, K. Winer and D. Weaire, Phys. Rev. Lett. **54**, 1392 (1985).

[156] M. Abramowitz and I. A. Stegun eds., *Handbook of Mathematical Functions* (Dover, New York, 1965), Ch. 24.

[157] K. P. Bogart, *Introductory Combinatorics* (Academic Press, New York, 1990).

[158] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Third Edition (John Wiley and Sons, New York, 1966).

[159] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Fifth Edition (Academia Press, New York, 1994).

Appendix A

The Geometry of Strained Diamond and Graphite

> My attempts to establish a universally acceptable notation were not very successful. In the course of time I was compelled to change my own notation, and thus I cannot complain that others felt the same need.
>
> – Max Born [70]

In this appendix we compute changes in bond lengths and angles, accurate to second order in strain, for the diamond and graphitic crystal structures (and, of course, we use our own unconventional notation). These results supply geometrical information needed in computing analytic formulae for elastic constants.

A.1 Diamond Lattice

A.1.1 First Neighbors

To be specific about the displacement of each atom under strain, center a Cartesian coordinate system on a representative atom at the origin, and number the (tetrahedral) first neighbors as follows,

$$\begin{aligned} \vec{r}_1 = & \tfrac{a_d}{4}(1,1,1)\,, & \vec{r}_2 &= \frac{a_d}{4}(-1,-1,1) \\ \vec{r}_3 = & \tfrac{a_d}{4}(-1,1,-1)\,, & \vec{r}_4 &= \frac{a_d}{4}(1,-1,-1), \end{aligned}$$

where a_d is the lattice constant. Denote the common first neighbor bond length by $r = \sqrt{3}a_d/4$. The volume per atom is $a_d^3/8$.

Appealing to cubic lattice symmetry, we thankfully need only consider the strain components $\varepsilon_1 = \varepsilon_{xx}$, $\varepsilon_2 = \varepsilon_{yy}$ and $\gamma_6 = 2\varepsilon_{xy}$ [75]. (Since $c_{66} = c_{44}$, γ_6 has the same effect on the energy as γ_4). In terms of these dimensionless strains, the distorted bond lengths are

$$\frac{\Delta r_i}{r} = \frac{1}{3}(\varepsilon_1 + \varepsilon_2) \pm \frac{1}{3}\gamma_6 + \frac{1}{9}(\varepsilon_1^2 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2) + \frac{1}{36}\gamma_6^2, \tag{A.1}$$

accurate to second order in the strain, where the upper sign applies to $i = 1,2$ and the lower to $i = 3,4$. At leading order, the uniaxial strains lengthen all four bonds, while the shear strain lengthens two and shortens the other two (consistent with volume conservation). The distorted angles are,

$$\Delta l_{12,34} = \frac{1}{9}\left[-4\left(\varepsilon_1 + \varepsilon_2 \pm \gamma_6\right) + \frac{2}{3}\left(\varepsilon_1^2 + \varepsilon_2^2 + 8\varepsilon_1\varepsilon_2\right) + \frac{5}{3}\gamma_6^2\right], \tag{A.2}$$

$$\Delta l_{13,24} = \frac{1}{9}\left[-4\varepsilon_1 + 8\varepsilon_2 + \frac{1}{3}\left(2\varepsilon_1^2 - 4\varepsilon_2^2 - 8\varepsilon_1\varepsilon_2\right) - \frac{1}{6}\gamma_6^2\right], \tag{A.3}$$

$$\Delta l_{14,23} = \frac{1}{9}\left[8\varepsilon_1 - 4\varepsilon_2 + \frac{1}{3}\left(-4\varepsilon_1^2 + 2\varepsilon_2^2 - 8\varepsilon_1\varepsilon_2\right) - \frac{1}{6}\gamma_6^2\right], \tag{A.4}$$

where in the first expression the upper sign refers to atoms 1 and 2 and lower to atoms 3 and 4. At first order, uniaxial tension opens four angles ($\Delta l < 0$) and closes two ($\Delta l > 0$), while the shear strain opens the angle between atoms 1 and 2 along the symmetry axis, closes the angle between atoms 3 and 4 perpendicular to the axis, and does not change the other four cross angles.

Using Eqs. (A.1) - (A.4), we can evaluate the expressions that arise in Taylor expansion of interatomic potentials to second order in the independent strains:

$$\sum_{i=1}^{4} \frac{\Delta r_i}{r} = \frac{4}{9}(\varepsilon_1^2 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2) + \frac{1}{9}\gamma_6^2, \tag{A.5}$$

$$\sum_{i=1}^{4} \left(\frac{\Delta r_i}{r}\right)^2 = \frac{4}{9}\left[(\varepsilon_1 + \varepsilon_2)^2 + \gamma_6^2\right], \tag{A.6}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \left(\frac{\Delta r_i}{r}\right)\left(\frac{\Delta r_j}{r}\right) = \frac{2}{9}[3(\varepsilon_1 + \varepsilon_2)^2 - \gamma_6^2] \tag{A.7}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \left(\frac{\Delta r_i}{r} + \frac{\Delta r_j}{r}\right) = \frac{4}{3}(\varepsilon_1^2 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2) + \frac{1}{3}\gamma_6^2, \tag{A.8}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \left[\left(\frac{\Delta r_i}{r}\right)^2 + \left(\frac{\Delta r_j}{r}\right)^2\right] = \frac{4}{3}\left[(\varepsilon_1 + \varepsilon_2)^2 + \gamma_6^2\right] \tag{A.9}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \Delta l_{ij} = \frac{8}{27}\gamma_6^2, \tag{A.10}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \Delta l_{ij}^2 = \frac{32}{81}\left[6\left(\varepsilon_1^2 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2\right) + \gamma_6^2\right], \tag{A.11}$$

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} \Delta l_{ij}\left(\frac{\Delta r_i}{r} + \frac{\Delta r_j}{r}\right) = -\frac{16}{27}\gamma_6^2. \tag{A.12}$$

A.1.2 Internal Relaxation

For C_{44} (or equivalently C_{66} here) we must allow internal relaxation of the interpenetrating FCC lattices. By symmetry the relaxation can only occur by moving the central atom at the origin along the z axis to the point $(0,0,z)$. Following Kleinman [81] and Harrison [76], we express the relaxation distance in terms of a parameter ζ via, $z = a\gamma_6\zeta/4$. With internal relaxation, the first neighbor bond lengths for the shear strain are,

$$\frac{\Delta r_i}{r} = \pm\frac{1}{3}(1-\zeta)\gamma_6 + \frac{1}{36}\left(3 + 6\zeta^2 - 2(1-\zeta)^2\right)\gamma_6^2, \tag{A.13}$$

where the upper sign refers to $i = 1,2$ and the lower to $i = 3,4$. If $\zeta = 1$, the first order change in all bond lengths vanishes. For simplicity, we only compute strained first neighbor bond angles to first order with internal relaxation. In that case,

$$\Delta l_{12} = -\Delta l_{34} = -\frac{4}{9}(1+2\zeta)\gamma_6, \tag{A.14}$$

and all other first order angular variations vanish.

A.1.3 Second Neighbors

The second neighbors in the diamond structure are the same as 12 first neighbors in an FCC crystal with the same lattice constant, *i. e.* $R(\pm1,\pm1,0)$, $R(\pm1,0,\pm1)$ and $R(0,\pm1,\pm1)$, where $R = a_d/\sqrt{2}$ is the second neighbor distance. The strained bond lengths are given by,

$$\frac{\Delta R}{R} = \frac{1}{2}(\varepsilon_1+\varepsilon_2) + \frac{1}{8}(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2) \pm \frac{1}{2}\gamma_6, \tag{A.15}$$

for $(1,1,0), (-1,-1,0)$ (upper sign) and $(1,-1,0), (-1,1,0)$ (lower sign),

$$\frac{\Delta R}{R} = \frac{1}{2}\varepsilon_1 + \frac{1}{8}\varepsilon_1^2 + \frac{1}{16}\gamma_6^2, \tag{A.16}$$

for $(\pm1,0,\pm1)$, and

$$\frac{\Delta R}{R} = \frac{1}{2}\varepsilon_2 + \frac{1}{8}\varepsilon_2^2 + \frac{1}{16}\gamma_6^2, \tag{A.17}$$

for $(0,\pm1,\pm1)$. The contributions from second neighbors to the expressions needed for (pair potential) elastic formulae are:

$$\sum_{i=1}^{12}\left(\frac{\Delta R_i}{R}\right) = \varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2+\frac{1}{2}\gamma_6^2, \tag{A.18}$$

$$\sum_{i=1}^{12}\left(\frac{\Delta R_i}{R}\right)^2 = 2(\varepsilon_1^2+\varepsilon_2^2+\varepsilon_1\varepsilon_2)+\gamma_6^2. \tag{A.19}$$

We do not consider angles involving second neighbors.

A.2 Graphitic Lattice

Now we repeat the same exercise for the graphitic structure. Here we only consider first neighbors, and hence restrict ourselves to covalent bonds within a single hexagonal plane. Relative to an atom at the origin, the first neighbors (all in the $z = 0$ plane) are

$$\vec{r}_1 = r(1,0),\ \vec{r}_2 = r\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right),\ \vec{r}_3 = r\left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right), \tag{A.20}$$

where $r = a_h/\sqrt{3}$ is the bond length and a_h is the in-plane lattice constant of the hexagonal array. The area per atom is $a_h^2\sqrt{3}/4$. We only consider the in-plane tensile strains ε_1 and ε_2. The γ_6 in-plane shear-strain dependence is related to ε_1 and ε_2 through $C_{66} = (C_{11} - C_{12})/2$ [85]. The deformed bond lengths and angles are

$$\Delta r_1/r = \varepsilon_1 + \frac{1}{2}\gamma_4^2, \tag{A.21}$$

$$\frac{\Delta r_i}{r} = \frac{1}{4}(\varepsilon_1 + 3\varepsilon_2) + \frac{3}{32}(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2) + \frac{1}{8}\gamma_4^2, \tag{A.22}$$

$$\Delta l_{1i} = \frac{3}{64}[8(\varepsilon_2 - \varepsilon_1) + 3\varepsilon_1^2 - 5\varepsilon_2^2 - 2\varepsilon_1\varepsilon_2] - \frac{3}{16}\gamma_4^2, \tag{A.23}$$

$$\Delta l_{23} = \frac{3}{4}(\varepsilon_1 - \varepsilon_2 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2) + \frac{3}{8}\gamma^2, \tag{A.24}$$

where $i = 2, 3$. The expressions needed for interatomic potential elastic constants are:

$$\sum_{i=1}^{3}\frac{\Delta r_i}{r} = \frac{3}{2}(\varepsilon_1 + \varepsilon_2) + \frac{3}{16}(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2) + \frac{3}{4}\gamma_4^2, \tag{A.25}$$

$$\sum_{i=1}^{3}\left(\frac{\Delta r_i}{r}\right)^2 = \frac{3}{8}\left[3(\varepsilon_1 + \varepsilon_2)^2 + 2\varepsilon_1\varepsilon_2\right], \tag{A.26}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\left(\frac{\Delta r_i}{r}\right)\left(\frac{\Delta r_j}{r}\right) = \frac{3}{16}[3(\varepsilon_1 + \varepsilon_2)^2 + 10\varepsilon_1\varepsilon_2], \tag{A.27}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\left(\frac{\Delta r_i}{r} + \frac{\Delta r_j}{r}\right) = 3(\varepsilon_1 + \varepsilon_2) + \frac{3}{8}(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2) + \frac{3}{2}\gamma_4^2, \tag{A.28}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\left[\left(\frac{\Delta r_i}{r}\right)^2 + \left(\frac{\Delta r_j}{r}\right)^2\right] = \frac{3}{4}\left[3(\varepsilon_1 + \varepsilon_2)^2 + 2\varepsilon_1\varepsilon_2\right], \tag{A.29}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\Delta l_{ij} = \frac{3}{32}[3(\varepsilon_1^2 + \varepsilon_2^2) - 10\varepsilon_1\varepsilon_2], \tag{A.30}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\Delta l_{ij}^{2} = \frac{27}{32}(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2), \tag{A.31}$$

$$\sum_{i=1}^{2}\sum_{j=i+1}^{3}\Delta l_{ij}\left(\frac{\Delta r_i}{r} + \frac{\Delta r_j}{r}\right) = -\frac{9}{16}(\varepsilon_1^1 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2). \tag{A.32}$$

The formulae in this section are somewhat tedious to compute and write down, but they are the starting point for the derivation of elastic constant formulae for wide classes of interatomic potentials.

Appendix B

Direct Inversion for Angular Forces

> The main reason behind the inability [of empirical potentials for silicon] to be more transferable is an inadequate description of angular forces.
>
> – H. Balamane, T. Halicioglu and W. A. Tiller [19]

Understanding the nature of angular forces in covalent solids has been a primary objective in this thesis. For small distortions from ideal configurations characterized by sp^3 and sp^2 hybrid covalent bonds, elastic constant analysis has been quite useful. For highly distorted and overcoordinated structures, inversion of cohesive energy curves is another fruitful method rooted in first principles. In the inversions presented in the main text, we have worked exclusively with cohesive energy versus volume curves. Although we have obtained angular functions by comparison of many such curves for different crystal structures, that approach only involves a small, discrete set of angles. Ideally, the angular function should somehow be extracted from shear strains or other reaction coordinates in which angles vary continuously with little change in volume.

In this appendix, we present mathematical proofs that this is at least a formal, if not also practical, possibility: within certain limitations, cohesive energy curves for angle-changing reaction coordinates can be inverted exactly to obtain the elusive angular dependence.

B.1 Formulation of the Problem

Our first task is to mathematically formulate the inversion problem in such a way that we might be able to solve it. We begin with the same assumptions about the functional form the interatomic potential as in Chapter 4, namely the separable, three-body cluster potential. This assumption, which precludes environment-dependence, shall be the main limitation on our results. As before we assume that the pair interaction is known (either by inversion or by assumption), so that the many-body portion of the energy, $F(\varepsilon)$, as a function of strain ε can be identified and expressed in terms of radial and angular functions,

$$F[g,h](\varepsilon) = \sum_i \sum_j \sum_{k>j} g(R_{ij}) g(R_{ik}) h(l_{ijk}), \tag{B.1}$$

where $l_{ijk} = \cos\theta_{ijk}$. A cohesive energy curve involves only one continuous degree of freedom, and hence is insufficient to determine more than one single-variable function in the interatomic potential. In Chapter 4, we use the idea of recursion to determine $g[h,F](r)$ with an assumed angular dependence in the case of uniform dilation. Here we show that the converse is also possible, to obtain $h[g,F](l)$ with an assumed radial function from shear strain data. In principle, a two-dimensional cohesive energy surface $F[g,h](\varepsilon_1, \varepsilon_2)$ could be inverted to simultaneously obtain both unknown functions $g[F](r)$ and $h[F](l)$, but since technical difficulties increase dramatically with the complexity of the attempted inversion, it is probably not worth pursuing such an ambitious course of action.

Let us distill the mathematics of Chapter 4 down to the central idea, which may be applied to the current problem: *recursively extend the solution from a point where the desired function is known to ones where it is not.* In the case of $g(r)$, we assume $\lim_{r\to\infty} g(r) = 0$, without loss of generality. This gives a starting point for recursive inversion, because in an infinitely expanded crystal, every bond is, of course, infinitely long. Inversion then proceeds by contracting the crystal to condensed volumes. In the case of $h(l)$, things are not so simple, but we can still make progress with reasonable assumptions.

We are fortunate that covalent structures have small coordinations, and hence far fewer angles than metallic structures. In the extreme cases of the diamond and graphitic structures, there is only one nearest neighbor bond angle. Elastic constant analysis shows that it is quite reasonable to assume $h(l) = 0$ in these cases and restrict the range of $g(r)$ to include only first neighbors, supplying two convenient starting points for inversion. Therefore, let $\varepsilon = 0$ correspond to the ideal diamond or graphitic lattice, and let l_o be the cosine of the first neighbor bond angle, $-1/3$ or $-1/2$, respectively. For finite strain ε, small enough not to alter the interaction topology (first neighbors), there will be a discrete set of M bond angle cosines. For each $l_m(\varepsilon)$, $m = 1, \ldots, M$, define the following quantities (analogous to α_{pq} in Section 4.2),

$$G_m(\varepsilon) = \sum_i \sum_j \sum_{k>j, l_{ijk}=l_m(\varepsilon)} g(R_{ij})g(R_{ik}), \tag{B.2}$$

where summation is over triplets forming the specified bond angle. The many-body energy then is expressed as,

$$F(\varepsilon) = \sum_{m=1}^{M} G_m(\varepsilon)h(l_m(\varepsilon)). \tag{B.3}$$

Note that it is not necessary for ε to be a homogeneous strain of the lattice for any of our arguments. In general, ε can represent any reaction coordinate that deforms the ideal crystal while maintaining invertability. Unfortunately, there is always a finite value of ε at which the inversion problem becomes singular.

B.2 The Constrained Case of an Even Angular Function

A complication is that a single strain will continuously sample the unknown function $h(l)$ in *two* directions away from l_o, corresponding to larger and smaller angles. (In the radial case, we start with infinite separation, and hence proceed in only one direction, namely smaller bond length.) As a first example, let us avoid this difficulty by artificially constraining the angular function to be even in the variable $\delta l = l - l_o$, *i.e.* $h(\delta l) = h(-\delta l)$. In that case the two independent directions collapse into one, and we have the recursion formula,

$$h(l_1(\varepsilon)) = \frac{1}{G_1}\left(F(\varepsilon) - \sum_{m=2}^{M} G_m(\varepsilon)h(l_m(\varepsilon))\right), \tag{B.4}$$

where the cosine deviations are numbered in order of decreasing magnitude, $|\delta l_1| > |\delta l_2| > \ldots > |\delta l_M|$ for each value of the strain. An explicit formula (without the unknown function on the right hand side) could in principle we obtained by recursive substitution, but in many cases it would quite cumbersome. Instead, we use a recursion procedure: start with $\varepsilon = 0$, where $h(l)$ is known, and solve for $h(\delta l_1(\varepsilon))$ in order of increasing ε.

The recursion formula above is easy to write down, but nontrivial to understand and apply, because issues of invertability and numerical stability are more subtle here than in the radial function case. In order for the inverse to exist, $l_1(\varepsilon)$ must be a strictly increasing function, so that at every stage $l_m(\varepsilon)$ is known for $m > 1$ from smaller values of ε. In general, it is not possible to find a reaction coordinate which opens the largest angle indefinitely, so there will be a singularity when $dl_1/d\varepsilon = 0$. The inversion may also become singular at a smaller value of ε if $l_1(\varepsilon)$ is discontinuous, which can occur when the interaction topology changes. For example, as ε is increased beyond a certain critical value ε_c, second neighbors may enter the interaction range (defined by the assumed function $g(r)$) and suddenly introduce new angles outside the range $0 \le \delta l < \delta l_1(\varepsilon)$, where $h(l)$ is known from $\varepsilon < \varepsilon_c$. Specifically, invertability is destroyed if $\lim_{\varepsilon \to \varepsilon_c^+} l_2 > \lim_{\varepsilon \to \varepsilon_c^-} l_1$, because we are then left with one equation and two unknowns (l_1 and l_2) at $\varepsilon = \varepsilon_c$.

The main source of numerical instability here (analogous to closely spaced first neighbor shells in the radial function case) is $(\delta l_1 - \delta l_2)/\delta l_1 \ll 1$. It is possible for this quantity to vanish for certain reaction coordinates at a nonzero value δl, but usually not with homogeneous stains. The universal problem with this instability is in starting the inversion procedure. At first all angular deviations are zero, and hence the right hand side of Eq. B.4 vanishes. Thus, the starting point for the recursion, $\varepsilon = 0$, is in fact a singularity! The trick to avoid the singularity and eliminate the associated instability is to use the analytic results of Chapter 3. Using the elastic constant formulae, it is straightforward to derive the curvature of the inverted angular function near the minimum that is appropriate for a given initial strain. Therefore, we can start the recursion at a small, positive value of ε and assume a known, parabolic form for $h(l)$ near the minimum, which should not present any problems. (This is equivalent to starting the radial function inversion at a large but finite first neighbor distance with an assumed tail rather than at infinite separation, where the radial function is actually known to vanish.)

B.3 The General Case of a Skewed Angular Function

The preceding case serves as an illustration containing most of the technical difficulties of the general case and giving the basic idea. The only idea needed to remove the (surely unreasonable) assumption of an even angular function is to consider *two* reaction coordinates, ε_1 and ε_2, in order to determine the two branches of the angular function for $\delta l > 0$ and $\delta l < 0$. Of course, each reaction path involves angles from both branches, so it will be necessary to invert both strain energy curves, $F_1(\varepsilon_1)$ and $F_2(\varepsilon_2)$, simultaneously. For the inverse to exist, however, the two reaction paths must be carefully chosen. We now discuss two general classes of reaction path pairs that lead to (formally) tractable inversion schemes.

B.3.1 Volume-Scaled Reaction Pairs

Once one useful reaction path ε_1 has been chosen, a second path ensuring invertability of the pair can be obtained by scaling the volume of the first path. Thus, the configuration corresponding to ε_2 on the second reaction path, has exactly the same structure (mainly, the same angles) as $\varepsilon_1 = \varepsilon_2$ on the first path, but with a different overall volume (which scales every bond length equally). The ratio of the two volumes may vary with ε_1, but may also be fixed for the entire reaction path. For example, we could consider simple shears strains $\varepsilon_1 = \varepsilon_2 = \gamma_4$ of the diamond lattice at two different volumes, corresponding to unstrained first neighbor distances $r_1 = 2.35$ Å and $r_2 = 2.55$ Å.

With such a "volume-scaled reaction pair", the inversion for the angular function proceeds as follows. For a particular value of $\varepsilon_1 = \varepsilon_2 = \varepsilon$, there is a set of M angles (the same in each of the two structures) whose cosines we number in decreasing order, $l_1 > l_2 > \ldots > l_M$. At a given stage in the inversion, the unknown variables are $h(l_1(\varepsilon))$ and $h(l_M(\varepsilon))$, while $h(l)$ is known for the range in between, $l_M(\varepsilon) < l < l_1(\varepsilon)$. In general, compared to the ideal angle l_o, both the positive and negative branches of $h(l)$ will be determined, since $\delta l_1 > 0$ and $\delta l_M < 0$. As before, define the radial quantities $G_{nm}(\varepsilon)$ via Eq. (B.2), for the two strain paths $n = 1, 2$, which again differ only in the overall volume. The recursion formula is a 2×2 linear system in this case,

$$G_{n1} h_1 + G_{nM} h_M = F_n - \sum_{m=2}^{M-1} G_{nm} h_m, \quad n = 1, 2, \tag{B.5}$$

where we use the shorthand notation, $h_m = h(l_m(\varepsilon))$, $G_{nm} = G_{nm}(\varepsilon)$ and $F_n = F_n(\varepsilon)$. If the volume difference between the two reaction paths is small, the determinant of the coefficient matrix, $|\mathrm{G}| = G_{11} G_{2M} - G_{21} G_{1M}$, at leading order is a linear combination of first derivatives of $g(r)$. These quantities can be small, thus causing instability, $|\mathrm{G}| \approx 0$. However, if the assumed $g(r)$ is not too flat, it may be possible to control the instability by choosing a large enough volume ratio.

B.3.2 Opening-Closing Reaction Pairs

Another approach involves choosing a pair of structurally different reaction paths such that for every state $(\varepsilon_1, \varepsilon_2)$ the largest angle among both structures is in one while the smallest angle is in the other. In this case one reaction path determines the largest angle (greatest opening of the ideal angle) and the other determines the smallest (greatest closing ideal angle). An example of such an "opening-closing reaction pair" might be a uniaxial strain taken in opposite directions, $\varepsilon_2 \propto -\varepsilon_1$ (compression and tension). For an opening-closing pair, in some sense, the matrix G of the previous section becomes diagonal. Although opening-closing pairs should provide greater numerical stability than volume-scaled pairs, the former may be difficult to construct, because a crystal deformation that produces very large angles (openings) typically also produces very small angles (closings).

To design an inversion procedure, couple the two reaction coordinates with a single parameter, $\varepsilon_1 = \varepsilon$ and $\varepsilon_2 = \alpha(\varepsilon)\varepsilon$. In the simplest case, take $\alpha(\varepsilon) = 1$, but the coupling function can be tuned to extend invertability or control numerical instability. Next independently number the cosines in the two structures, $\{l_{1m}(\varepsilon)\}$ and $\{l_{2m}(\varepsilon)\}$, in decreasing order, $l_{n1} > l_{n2} > \ldots > l_{nM_n}$. In general, the number of angles in each structure may be different, $M_1 \neq M_2$. The opening-closing condition requires that the largest cosine overall is in the first structure, $l_{max}(\varepsilon) = l_{11}(\varepsilon)$, and the smallest is in the other, $l_{min}(\varepsilon) = l_{2M_2}(\varepsilon)$. With these definitions, the recursion formulae for the unknown function $h(l)$ are,

$$h_{max} = \frac{1}{G_{11}} \left(F_1 - \sum_{m=2}^{M_1} G_{1m} h_{1m} \right), \tag{B.6}$$

$$h_{min} = \frac{1}{G_{2M_2}} \left(F_2 - \sum_{m=1}^{M_2-1} G_{2m} h_{2m} \right). \tag{B.7}$$

In general, the inverse will exist as long as $l_{max}(\varepsilon)$ and $l_{min}(\varepsilon)$ are continuous monotonic functions (increasing and decreasing, respectively). As before, the procedure must be started with parabolic forms for $h_{max}(\varepsilon)$ and $h_{min}(\varepsilon)$, derived analytically (*e.g.* from elastic constant formulae).

B.4 Conclusion

One can formally invert (pairs of) cohesive energy curves for carefully chosen reaction coordinates to obtain angular functions, $h[g, F](l)$. The technical complications related to invertability and numerical stability are more subtle than in radial function case, so it is not clear that direct angular inversions will succeed in practice. If meaningful inversions are possible, one might envision iterating between radial and angular inversions or selecting an optimal radial function that causes collapse of inverted angular functions, by analogy to Section 4.3. However, such inversions are likely to become so complex that no advantage would be gained over the usual fitting approach.

A better use of these formulae would be in testing the assumptions of hypothetical models of bonding like EDIP directly against *ab initio* data. Although these inversions may be challenging to perform, the results are desperately needed. For example, by comparing inverted angular functions from strains of the graphitic and diamond lattices, we could test the EDIP conjecture that the shape of the angular function $H(x)$ is the same for $Z = 3$ and $Z = 4$. We could also see to what extent a single angular function can handle diverse reaction paths at fixed coordination. These assumptions are surely imperfect, but it is not clear *a priori* how bad they will turn out to be. Perhaps, as in the case of the Harrison model for unrelaxed elasticity, the EDIP model will be fortuitously good. Alternatively, the results may be so bad that different hypotheses are required. In any case, producing meaningful *ab initio* angular functions without any *ad hoc* fitting would be a welcome theoretical advance.

Appendix C

Recursion and the Möbius Inversion Formula

> The belief that pure mathematics is only fortuitously useful is widely shared, even by mathematicians.
>
> – John Maddux [96]

The Möbius inversion formula from number theory [94, 95] has gained attention in the current physics literature through the work of Chen, who has proved a generalization of the theorem to continuous variables [93]. Although its practical utility remains to be seen, the Chen-Möbius theorem provides an elegant formal method to solve nonlinear inversion problems. As described in Chapter 4.1, an important application in theory of solid cohesion is the problem of inverting a crystal energy versus volume curve to obtain the interatomic forces (assumed to act radially between pairs of atoms) [97]. In that case, the Chen-Möbius theorem is a more compact but equivalent statement of an inversion formula originally derived by Carlsson, Gelatt and Ehrenreich (CGE) [90]. In Chapter 4.1.4, a simple, recursive proof of the CGE formula is given that provides insight into the physical meaning of the inversion process and generalizes it to more complicated situations (nonradial forces and nonuniform deformations). In this Appendix, we shall see that the idea of recursion can also be used to derive a discrete analog of the CGE inversion formula. By comparing with the Möbius inversion formula, an interesting expression for the number-theoretic Möbius function can also be derived in terms of combinatorial quantities that generalize the Stirling numbers of the second kind [157].

C.1 A Recursive Approach to Möbius Inversion

We begin by stating the Möbius inversion formula.

Theorem: Given $f : N \to N$, define $F : N \to N$ by

$$F(n) = \sum_{d|n} f(d). \tag{C.1}$$

Then $\forall n \in N$,

$$f(n) = \sum_{d|n} \mu(d) F(n/d), \tag{C.2}$$

where $\mu : N \to \{0, 1, -1\}$ is the Möbius function,

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ (-1)^\ell & \text{if n = } \text{ product of } \ell \text{ distinct primes} \\ 0 & \text{otherwise} \end{cases} . \tag{C.3}$$

Proof: Substitute Eq. (C.1) in the right side of Eq. (C.2) and simplify:

$$\sum_{d|n} \mu(d) F(n/d) = \sum_{d|n} \mu(d) \sum_{c|(n/d)} f(c) = \sum_{c|n} f(c) \sum_{d|(n/c)} \mu(d) = f(n) \tag{C.4}$$

using $\sum_{d|n} \mu(d) = \delta_{n1}$ in the last step [95].

Following the arguments of Section 4.1.4, we can easily derive another, albeit more complicated, inversion formula.

Lemma: Let s be the number of factors (not necessarily distinct) in the prime decomposition of n. Then, given the assumptions of the theorem, $\forall n \in N$,

$$\begin{aligned} f(n) &= F(n) - \sum_{d_1|n}{}' F(n/d_1) + \sum_{d_1|n}{}' \sum_{d_2|n}{}' F(n/d_1 d_2) \\ &- \sum_{d_1|n}{}' \sum_{d_2|n}{}' \sum_{d_3|n}{}' F(n/d_1 d_2 d_3) + \ldots (-1)^s F(1), \end{aligned} \tag{C.5}$$

where the kth multiple sum is restricted so that $(d_1 d_2 \ldots d_k)|n$ and $\sum_d'$ denotes summation with $d = 1$ excluded.

Proof: Rewrite Eq. (C.1) as,

$$F(n) = \sum_{d|n} f(n/d), \tag{C.6}$$

and solve for the $d = 1$ term of the sum,

$$f(n) = F(n) - \sum_{d|n}{}' f(n/d). \tag{C.7}$$

The result follows by recursive substitution.

In spite of the complexity of the explicit formula, recursion clearly reveals how the process of Möbius inversion takes place, in analogy with the contraction of a greatly expanded crystal down to condensed volumes in the case of solid cohesion. Let us order the divisors of n in decreasing order, $n = d^{(1)} > d^{(2)} > \ldots > d^{(s)} = 1$. The procedure starts with $F(n)$ as an initial guess for the unknown $f(n)$. The error is corrected by subtracting $f(d^{(i)})$ for all *smaller* divisors, $i \geq 2$. After one iteration, we have

$$f(n) = F(n) - \sum_{d_1|n}{}' F(n/d_1) + \sum_{d_1|n}{}' \sum_{d_2|n}{}' f(n/d_1 d_2). \tag{C.8}$$

Now the same initial guess, $f(d^{(i)}) = F(d^{(i)})$, has been applied to get the first correction from the $i \geq 2$ divisors. The error at this stage is removed by adding $f(d^{(i)})$ for smaller divisors, $i \geq 3$, and $F(n)$ and $F(d^{(2)})$ have already made their full contributions. As the process continues, new corrections come from successively smaller divisors until the last step, when the entire error is concentrated in $F(1)$. Note that the convergence of the (finite) series in Eq. (C.5) is quite slow, as pointed out by Chen and Ren in the case of the CGE formula [97], because the terms are oscillatory and often quite large, as described in the next section.

C.2 A Combinatorial Expression for the Möbius Function

By combining like terms, Eq. (C.5) can be written in the form,

$$f(n) = \sum_{d|n} c(d) F(n/d). \tag{C.9}$$

where, of course, $c(d) = \mu(d)$ by linear independence. In this way, the Lemma can be used to prove the following result.

Theorem: Let d have prime decomposition $d = \prod_{i=1}^{\ell} p_i^{\alpha_i}$, with $m = \sum_{i=1}^{\ell} \alpha_i$. Then,

$$\mu(d) = \sum_{k=1}^{m} (-1)^k \; \mathcal{B}(d,k), \tag{C.10}$$

where $\mathcal{B}(d,k)$ is the number of distinct decompositions of d into k nontrivial factors.

Equivalently, $\mathcal{B}(d,k) = \mathcal{B}(\{\alpha_i\}_{i=1}^{\ell}, k)$ is the solution to the following counting problem[1]:

Exercise: Suppose we have a set of m colored balls with α_i identical balls of the ith color, $i = 1, \ldots, \ell$. How many ways are there to distribute the balls among k distinct containers, placing at least one ball in each?

The general solution of the exercise is quite difficult, perhaps even intractable, so we pose it as a challenge to the reader. Whatever the solution, however, it must satisfy the sum rule of Eq. (C.10), indicating a subtle cancelation of these complex quantities during recursive inversion. There are two limiting cases in which the exercise reduces to well-known (but nontrivial) counting problems:

[1] Analysis of the counting exercise was done in close collaboration with Adam Lupu-Sax.

1. In the case of indistinguishable balls of a single color ($\ell = 1$ and $m = \alpha_1$), which corresponds to $d = p^m$ for some prime p, a binomial coefficient, $\mathcal{B}(d, k) = \binom{m-1}{k-1}$, solves the exercise [157][2]. In this case, the sum rule implied by Eq. (C.10) is,

$$\sum_{k=1}^{m}(-1)^k \mathcal{B}(p^m, k) \;=\; \begin{cases} 1 & \text{if } m = 1 \\ -(1-1)^{m-1} = 0 & \text{if } m > 1 \end{cases} \;=\; \mu(p^m), \qquad \text{(C.11)}$$

which is easily verified with the binomial formula.

2. In the case of distinguishable balls, ($\alpha_i = 1, \forall i = 1, \ldots, \ell$), which corresponds to d being a product of ℓ distinct primes, the solution to the exercise is $\mathcal{B}(d, k) = k!\, \mathcal{S}_\ell^{(k)}$, where $\mathcal{S}_\ell^{(k)}$ is a Stirling number of the second kind [157]. In this case, as a simple consequence of our lemma, we have the sum rule for Stirling numbers of the second kind [156], which we state as a corollary.

Corollary:

$$\sum_{k=1}^{\ell}(-1)^k k!\, \mathcal{S}_\ell^{(k)} = (-1)^\ell. \qquad \text{(C.12)}$$

This same fact can also be derived using standard combinatorial formulae. Substituting the expression for Stirling numbers of the second kind [156, 157], we have

$$c(d) = \sum_{k=1}^{\ell}\sum_{j=0}^{k}(-1)^j \binom{k}{j} j^\ell. \qquad \text{(C.13)}$$

By exchanging the order of summation and performing the inner sum using Eq. (0.151.1) of Ref. [159], we obtain,

$$c(d) = \sum_{j=1}^{\ell}(-1)^j \binom{\ell+1}{j+1} j^\ell. \qquad \text{(C.14)}$$

[2]This formula is easily proved with the following insight [158]: Represent the balls by a sequence of m dots and the container walls by $k-1$ vertical lines. The number of ways to distribute the balls into the containers is equal to the number of ways to distribute the lines among the $m-1$ spaces between the dots, one per space and without regard to order, which is simply $\binom{m-1}{k-1}$. For example, $B(p^4, 3) = 3$: $\bullet|\bullet|\bullet\bullet$, $\bullet|\bullet\bullet|\bullet$, $\bullet\bullet|\bullet|\bullet$.

Next we separate the $j = \ell$ term and break the sum into two parts using Eq. (24.1.1.II.A) of Ref. [156],

$$c(d) = (-1)^{\ell}\ell^{\ell} + \sum_{j=1}^{\ell-1}(-1)^{j}\binom{\ell}{j}j^{\ell} - \sum_{j=1}^{\ell}(-1)^{j}\binom{\ell}{j}(-1+j)^{\ell}. \tag{C.15}$$

Finally, evaluating the sums with Eqs. (0.154.4) and (0.154.6) of Ref. [159], we arrive at the desired result, $c(d) = (-1)^{\ell}$.

The general case of m balls of an arbitrary number of distinct colors is much more complicated, but the answer must be related to the Möbius function through Eq. (C.10).

In conclusion, we use recent developments in the theory of solid cohesion to provide a fresh perspective on the process of Möbius inversion. We prove a variant of the Möbius inversion formula and from it derive a representation of the Möbius function in terms of the combinatorial quantities $\mathcal{B}(d,k)$, which reduce to the Stirling numbers of the second kind in a special case. In this way, we arrive at a unified view of a wide class of counting problems related to the Möbius function.