Theory and application of large-scale electronic structure calculations

Takeo Hoshi

Sep, 2003

This electronic file is based on the thesis of Takeo Hoshi, published at Sep. 2003, from School of engineering, University of Tokyo. See the following URL as the contact address : http://fujimac.t.u-tokyo.ac.jp/hoshi/

This PDF file is compiled at 31. Oct. 2003.

Contents

I Introduction and review **1**

1 Introduction **2**

2 Current electronic structure theory **5**
2.1 Density functional theory 6
2.2 Construction of tight-binding Hamiltonians 9
2.3 Generalized Wannier states 14
2.4 Fundamentals of order-N methods 17
2.5 Practical order-N methods 21

II Theory **30**

3 Tight-binding theory among elements and phases **31**
3.1 Universal tight-binding theory 32
3.2 Liquid and surface phases of silicon 36
3.3 Summary and discussion 41

4 Theories for large-scale calculations **43**
4.1 Quantum mechanics with one-body density matrix 44
4.2 Mean-field equation for generalized Wannier states 50
4.3 Variational and perturbative order-N methods 56
4.4 Hybrid scheme by dividing Hilbert space 60
4.5 Summary and future aspects 67

5 Details and applications **69**
5.1 Perturbative order-N method 70
5.2 Variational order-N method 73
5.3 Hybrid scheme 76
5.4 Parallelization of perturbative order-N method 80
5.5 Wannier states in diamond structure solids 82

III Application to fracture of nanocrystalline silicon 96

6 Backgrounds 97
6.1 Fracture theory and silicon 98
6.2 Requirement on atomistic theory for fracture 101
6.3 Energetics of Si(001) surface 103

7 Fracture of nanocrystalline silicon 110
7.1 Purpose 111
7.2 Basic properties of fracture simulation 114
7.3 Effect of dehybridization mechanism 122
7.4 Technical details of the dynamical simulation 130
7.5 Fracture simulation with thousands of atoms 138
7.6 Fracture simulation with 10^4-10^5 atoms 150
7.7 Summary and discussion 165

8 Summary and general discussion 167

Appendices

A Note on electronic structure theory 171
A.1 First-principle molecular dynamics and limitation of LDA 172
A.2 Tight-binding formulation 175

B Theory of elasticity 180
B.1 Theory in cubic symmetry 181
B.2 Simple classical model in tetrahedral structure 192
B.3 Theory in isotropic medium 201

C Continuum theory of fracture 208
C.1 Theory of elasticity in isotropic 2D medium 209
C.2 Theories of fracture 217

D Miscellaneous notes 221
D.1 Conventional Wannier state in one-dimensional system 222
D.2 Density matrix in free electron system 224
D.3 Lanczos method 226
D.4 Verlet algorithm in molecular dynamics 231

Acknowledgements 234

Bibliography 235

Part I

Introduction and review

Chapter 1

Introduction

Nanoscale materials are directly governed by quantum mechanical freedoms of electron systems. Nowadays, electron systems of realistic materials are treated by the *ab initio* electronic structure calculations that are based on the density functional theory [1, 2], the first-principle molecular dynamics [3], and related theories developed for decades. A typical system size of present *ab initio* calculations is, however, on the order of 10^2 atoms and new practical theories are required for nanoscale calculations. This thesis is devoted to the theory of large-scale electronic structure calculations and its application to realistic materials.

A keyword of the present thesis is 'order-N'. Since a standard quantum mechanical calculation is reduced to the matrix diagonalization procedure, its computational cost is proportional to N^3, where N is the system size. This fact restricts severely the system size of quantum mechanical calculations. The term order-N method is a general name of electronic structure methods in which the computational cost is proportional to the system size. Figure 1.1 demonstrates our large-scale calculation among $10^2 - 10^6$ atoms. The resultant CPU time in the order-N method is almost ideally proportional to the system size up to over one million atoms, while the exact diagonalization method shows an $O(N^3)$ scaling property in the CPU time. The present system sizes ($10^2 - 10^6$ atoms) are directly related to the present or next generation technologies. For example, the circuit design rule of the present processors, such as Pentium 4^{TM}, is based on the length scale of 10^2 nm, which corresponds to about 200 atomic layers. The design rule in finer scales is now focused as an urgent industrial issue [4]. Here it should be emphasized that the large-scale calculations in the present thesis are done, not by a novel hardware environment, such as a parallel computer, but by a novel theory of electronic structure calculations. We should say, however, that such large-scale calculations are possible within a limited applicability and/or at the sacrifice of accuracy. Therefore, the methods for a nanoscale material simulation should be properly constructed, in system size, applicability and accuracy, according to its purpose.

Now we discuss the nanoscale material theory from the viewpoint of simulation methods. The present simulations of material structures can be classified into three categories; The first category contains methods based on quantum mechanics of electronic structures, such as the first-principle molecular dynamics. The second category contains methods based on classical mechanics, such as classical molecular dynamics. The last category contains methods based on continuum mechanics, such

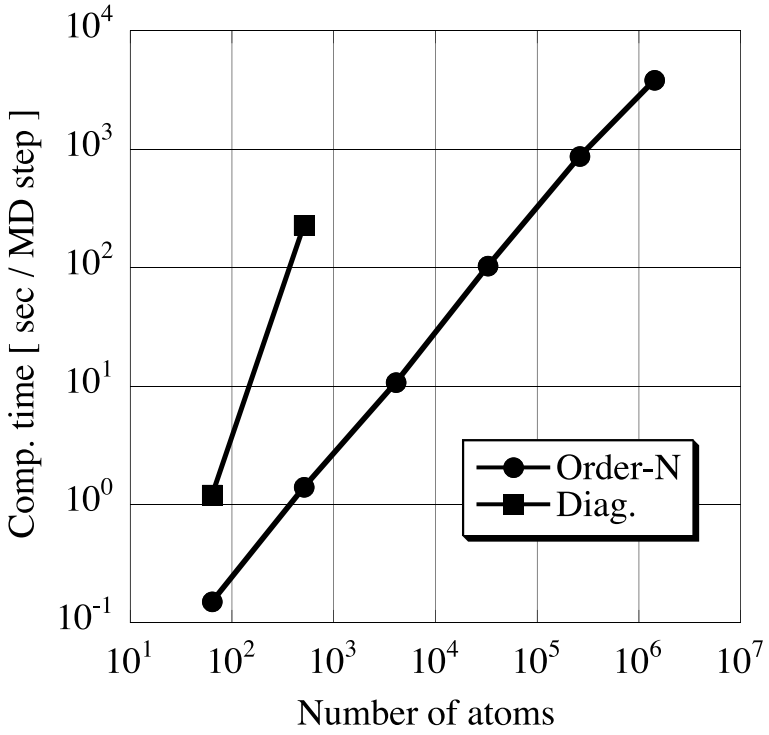


Figure 1.1: The computational time of bulk silicon as the function of the number of atoms (N), up to 1,423,909 atoms [5]; The CPU time is measured for one time step in the molecular dynamics (MD) simulation. A tight-binding Hamiltonian is solved using the exact diagonalization method and the perturbative 'order-N' method (See Section 4.3). We use a standard work station with one Pentium 4$^{\rm TM}$ processor and 2 GB of RAM.

as finite element method. Conventionally, the targets of the three categories are divided in accordance with the system size. Recently, several methodologies are focused that bridge the above three categories. Such methodologies are often called 'multiscale mechanics'. In this context, the present system size (10^2-10^6 atoms) is the intermediate size between those in the quantum and classical mechanical methods.

The purpose of the present thesis, however, is not only to construct practical simulation methods for the above intermediate system size, but also to construct a guiding principle for bridging the three principles of mechanics. The key concept is the total energy functional. The total energy functional is well defined among the above three principles of mechanics in the sense that a material structure is determined by minimizing the total energy functional.

The present thesis is mainly devoted to theories for the simplification of the *ab initio* total energy functional with respect to quantum mechanical freedoms. The following three concepts will be discussed by simplifying the total energy functional; The first one is the tight-binding formulation of the electronic structure energy, particularly its universality. The universality of the tight-binding Hamiltonian is justified from the *ab initio* theory and gives a systematic investigation among different elements and phases. The second one is the order-N methods. Particularly, we derive two practical order-N algorithms, variational one and perturbative one,

based on the generalized Wannier states. The third one is the hybrid scheme by dividing the occupied Hilbert space. The division is done with respect to the one-body density matrix, which is well defined in quantum mechanics. As an important application of the large-scale calculations, we discuss the dynamical brittle fracture of nanocrystalline silicon.

This thesis is organized as follows; Part I is devoted to the introduction (this chapter) and a review of the current electronic structure theory (Chapter 2). In Part II, we construct novel theories for large-scale calculations in Chapter 3 and Chapter 4. Several exact quantum mechanical equations are derived as the foundation of large-scale calculations. The technical details and some applications of the theories are described in Chapter 5. Part III is devoted to the application to the fracture dynamics of nanocrystalline silicon, particularly its possible differences from that of macroscale samples. In Chapter 6, we explain the background of the fracture simulations. In Chapter 7, the fracture dynamics is simulated with up to 10^5 atoms. The analysis of the results shows the crucial role of the quantum mechanical freedoms in electronic structures. In the last chapter, Chapter 8, we describe the summary and general discussions. Several appendices are also prepared.

The atomic unit ($m_{\rm e} = \hbar = |e| = 1$) is used throughout the present thesis, except where indicated. As a typical transferable tight-binding Hamiltonian for silicon, we use one in Ref. [6], except where indicated.

Chapter 2

Current electronic structure theory

2.1 Density functional theory

Here we gives a brief introduction to the *ab initio* electronic structure theory. Especially we focus on the density functional theory (DFT) within the local density approximation (LDA) as the standard method for condensed matters. The density functional theory is based on several fundamental theories, such as the Hohenberg-Kohn theorem [1], the Kohn-Sham equation [2], and the Janak theorem [7]. See reviews, such as Ref. [8]. Here we explain only the resultant formulations that are used in the present standard calculations.

Formulation

Suppose an electronic system in an external potential $V_{\rm ext}(\boldsymbol{r})$. The electronic system described by a single Slater determinant of occupied one-electron wave functions $\{\phi_i\}_i$. In a realistic system of many atoms, the external potential $V_{\rm ext}$ corresponds to the sum of the Coulomb potentials from the nucleus of the atoms. Hereafter we limit the discussion to the para-magnetic cases. The charge density is defined by

$$n(\boldsymbol{r}) = \sum_i^{\rm occ.} |\phi_i(\boldsymbol{r})|^2, \tag{2.1}$$

and the total energy is given by

$$E_{\rm tot} \equiv E_{\rm kin} + E_{\rm ext} + E_{\rm H} + E_{\rm XC}, \tag{2.2}$$

$$E_{\rm kin} = \sum_i \int d\boldsymbol{r} \phi_i^*(\boldsymbol{r}) \frac{-\nabla^2}{2} \phi_i(\boldsymbol{r}), \tag{2.3}$$

$$E_{\rm ext} = \int V_{\rm ext}(\boldsymbol{r}) n(\boldsymbol{r}) d\boldsymbol{r}, \tag{2.4}$$

$$E_{\rm H} = \frac{1}{2} \int\int \frac{n(\boldsymbol{r}) n(\boldsymbol{r}') d\boldsymbol{r} d\boldsymbol{r}'}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{2.5}$$

Here $E_{\rm kin}, E_{\rm ext}, E_{\rm H}, E_{\rm XC}$ are the kinetic energy, the external potential energy, the Hartree energy and the Exchange-correlation energy, respectively. The exchange-correlation energy $E_{\rm XC} = E_{\rm XC}[n]$ is a given functional of the charge density $n(\boldsymbol{r})$. The present standard functional is that within the local density approximation (LDA)

$$E_{\rm XC} = E_{\rm XC}^{\rm (LDA)} \equiv \int n(\boldsymbol{r}) \varepsilon_{\rm XC}^{\rm (LDA)}(n(\boldsymbol{r})) d\boldsymbol{r}. \tag{2.6}$$

Here $\varepsilon_{\rm XC}^{\rm (LDA)}(n(\boldsymbol{r}))$ is a function, not functional, of the local charge density. The explicit form of the function $\varepsilon_{\rm XC}^{\rm (LDA)}(n)$ is determined so as to reproduce the exact ground-state energy of the homogeneous electron gas.

As in standard variational procedures, the equation for one wave function is given by

$$\frac{\delta}{\delta \phi_i^*} \left\{ E_{\rm tot} - \sum_{j,k}^{\rm occ.} \varepsilon_{kj} \langle \phi_k | \phi_j \rangle \right\} = 0, \tag{2.7}$$

where the matrix ε_{kj} is the Lagrange multiplier with respect to the orthogonality constraints between the wave functions

$$\langle\phi_i|\phi_j\rangle = \delta_{ij}. \tag{2.8}$$

The equation(2.7) is rewritten as

$$H_{\rm KS}|\phi_i\rangle - \sum_j^{\rm occ.} \varepsilon_{ij}|\phi_j\rangle = 0 \tag{2.9}$$

with the effective Hamiltonian of

$$H_{\rm KS} \equiv \frac{-\nabla^2}{2} + V_{\rm eff}(\boldsymbol{r}), \tag{2.10}$$

where

$$\begin{aligned} V_{\rm eff}(\boldsymbol{r}) &\equiv \frac{\delta E_{\rm ext}}{\delta n(\boldsymbol{r})} + \frac{\delta E_{\rm H}}{\delta n(\boldsymbol{r})} + \frac{\delta E_{\rm XC}}{\delta n(\boldsymbol{r})} \\ &= V_{\rm ext}(\boldsymbol{r}) + \int \frac{n(\boldsymbol{r}')d\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{\delta E_{\rm XC}}{\delta n(\boldsymbol{r})}. \end{aligned} \tag{2.11}$$

This Hamiltonian $H_{\rm KS}$ is called the Kohn-Sham Hamiltonian. The solution of Eqs. (2.8) and (2.9) gives the ground state within the single Slater determinants. Using Eqs.(2.8) and (2.9), one can find that the matrix ε_{ij} is Hermitian in the ground state ($\varepsilon^*_{ij} = \varepsilon_{ij}$) and is given as

$$\varepsilon_{ij} = \langle\phi_j|H_{\rm KS}|\phi_i\rangle. \tag{2.12}$$

Any physical quantity $\langle\hat{X}\rangle$

$$\langle\hat{X}\rangle \equiv \sum_k^{\rm occ}\langle\phi_k|\hat{X}|\phi_k\rangle \tag{2.13}$$

is invariant under the unitary transforms with respect to the occupied wave functions

$$|\phi_i\rangle \rightarrow |\phi'_i\rangle \equiv \sum_j^{\rm occ.} U_{ij}|\phi_j\rangle, \tag{2.14}$$

where U_{ij} is a unitary matrix. This freedom is called the 'unitary freedom' in the sense that the wave function has a freedom that does not affect any physical quantity. If one fixes the above unitary freedom so that the matrix ε_{ij} becomes diagonal ($\varepsilon_{ij} = \delta_{ij}\varepsilon_i^{\rm (eig)}$), Eq.(2.9) is reduced to an eigen value problem:

$$H_{\rm KS}|\phi_i^{\rm (eig)}\rangle = \varepsilon_i^{\rm (eig)}|\phi_i^{\rm (eig)}\rangle. \tag{2.15}$$

This equation is the Kohn-Sham equation and is solved in practical electronic-structure calculations.

Discussions

Now the one-body density matrix $\hat{\rho}$ is introduced as

$$\hat{\rho} \equiv \sum_{i}^{\text{occ.}} |\phi_i\rangle\langle\phi_i| \tag{2.16}$$

or

$$\rho(\boldsymbol{r}, \boldsymbol{r}') \equiv \sum_{i}^{\text{occ}} \phi_i^*(\boldsymbol{r}')\phi_i(\boldsymbol{r}). \tag{2.17}$$

The commutation relation

$$0 = \hat{H}_{\text{KS}}\hat{\rho} - \hat{\rho}\hat{H}_{\text{KS}} \tag{2.18}$$

is satisfied. The density matrix $\hat{\rho}$ is unique and invariant under the unitary transforms of Eq. (2.14). Any physical quantity $\langle\hat{X}\rangle$ is described, with the density matrix ρ, in the trace form of

$$\langle\hat{X}\rangle \equiv \sum_{i}^{\text{occ.}}\langle\phi_i|\hat{X}|\phi_i\rangle = \text{Tr}[\hat{\rho}\hat{X}] = \int d\boldsymbol{r}\int d\boldsymbol{r}'\rho(\boldsymbol{r}, \boldsymbol{r}')X(\boldsymbol{r}', \boldsymbol{r}). \tag{2.19}$$

Note that the DFT can be generalized by the fractional occupation formalism [9, 10, 11], in which the charge density is redefined as

$$n(\boldsymbol{r}) = \sum_{i} f_i|\phi_i(\boldsymbol{r})|^2. \tag{2.20}$$

In the fractional occupation formalism, the one body matrix should be redefined as

$$\hat{\rho} \equiv \sum_{i} f_i|\phi_i\rangle\langle\phi_i|, \tag{2.21}$$

in which the unitary freedom exists only among the electronic states $\{\phi_i\}$ whose occupations $\{f_i\}$ are the same value.

In this section, we have explained the foundations of the DFT. A couple of related topics will be discussed in Appendix A.1, that is, the first-principle molecular dynamics and the limitation of the LDA.

2.2 Construction of tight-binding Hamiltonians

Here we explain that the tight-binding Hamiltonian can be constructed from the *ab initio* theory. The construction is done systematically in the linear muffin tin orbital (LMTO) theory [12, 13]. In this section, we will discuss the theory within the atomic sphere approximation. The 'muffin tin' means spherical regions whose centers are located at atom sites. Radius for each atomic sphere is properly given so that the total volume of the spheres is equal to the volume of the system.

Electronic structure theory as scattering problem

Let the Kohn-Sham potential be transformed into its spherical average within each muffin tin region;

$$V_{\rm eff}(\boldsymbol{r}) \Rightarrow V_{\rm eff}(|\boldsymbol{r}|) \quad \text{at} \quad r < R, \tag{2.22}$$

where R is the radius of the corresponding spherical region. Within an atomic sphere ($r < R$), we can write the Kohn-Sham equation as

$$\left(-\frac{\nabla^2}{2} + V_{\rm eff}(r) - \varepsilon\right)|\phi_{lm}\rangle = 0 \tag{2.23}$$

with the suffices of the angular momentum (l, m). The above situation is similar to that in an isolated atom. A crucial difference between isolated atoms and condensed matters is the difference in the boundary condition. In isolated atoms, the vanishing boundary condition

$$|\phi_{lm}(\varepsilon)\rangle = 0 \quad (r \to \infty) \tag{2.24}$$

is imposed, which results in the quantization of the energy ($\varepsilon = \varepsilon_i$). In condensed matters, on the other hand, a continuum energy band is possible, due to the lack of the vanishing boundary condition. Electronic structures in condensed matters are based on the potential scattering problem. If the effective potential is supposed to be constant outside the sphere

$$V_{\rm eff}(r) = V_0 \quad \text{at} \quad r > R, \tag{2.25}$$

the problem is reduced to a potential scattering problem. The free electron system is a simple example, in which the Kohn-Sham equation is reduced to the Helmholtz equation

$$(\Delta + k^2)|\phi\rangle = 0, \tag{2.26}$$

where $k \equiv \sqrt{2\varepsilon}$ and $\varepsilon > 0$. The general solution of Eq. (2.26) is written as a linear combination of spherical waves

$$\phi(\boldsymbol{r}) = \sum_{lm} \{A_{lm} j_l(kr) + B_{lm} n_l(kr)\} Y_{lm}(\hat{\boldsymbol{r}}), \tag{2.27}$$

with the arbitrary constants A_{lm} and B_{lm}. The function $j_l(x)$ or $n_l(x)$ is the spherical Bessel or Neumann functions, respectively. A plane wave $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ is also a solution of Eq. (2.26) and is written by the spherical waves in

$$e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = e^{ikr\cos\theta} = \sum_{l=0}^{\infty}(2l+1) i^l j_l(kr) P_l(\cos\theta), \tag{2.28}$$

where the z axis is chosen in the direction of $\boldsymbol{k}$.

Energy linearization and LMTO theory

Now we are back to Eq. (2.23). One of the most important concepts for the current electronic structure theory is the concept called 'energy linearization'. The linearization concept is the foundation of, not only the LMTO theory [12, 13], but also the *ab initio* pseudo potential theory [14, 15, 16, 17]. For a given 'reference' energy $\varepsilon_{\rm ref}$, the wave function $|\phi_{lm}(\varepsilon)\rangle$ can be approximated within the linear expansion near the reference energy($\varepsilon \approx \varepsilon_{\rm ref}$);

$$|\phi_{lm}(\varepsilon)\rangle \approx |\phi_{lm}(\varepsilon_{\rm ref})\rangle + (\varepsilon - \varepsilon_{\rm ref})|\dot{\phi}_{lm}(\varepsilon_{\rm ref})\rangle \tag{2.29}$$

where $|\dot{\phi}_{lm}(\varepsilon)\rangle$ is the energy derivative of the wave function

$$|\dot{\phi}_{lm}(\varepsilon)\rangle \equiv \frac{d|\phi_{lm}(\varepsilon)\rangle}{d\varepsilon}. \tag{2.30}$$

The reference energy ε is chosen independently for each angular momentum. The reference energy should be, basically, chosen at the center of an energy band in condensed matters. The wave function ϕ is normalized in the spherical region

$$\langle\phi_{lm}|\phi_{lm}\rangle_R = 1 \tag{2.31}$$

where

$$\langle\cdots\rangle_R \equiv \int_{r<R} \cdots d\boldsymbol{r}. \tag{2.32}$$

Differentiating Eq. (2.31) with respect to the energy, we obtain the orthogonal relation between ϕ and $\dot{\phi}$

$$\langle\phi_{lm}|\dot{\phi}_{lm}\rangle_R = 0 \tag{2.33}$$

within the sphere. In the LMTO method, the functions

$$|\chi_{Ilm}\rangle \equiv |\phi_{Ilm}\rangle + \sum_{Jl'm'} h_{Ilm,Jl'm'}|\dot{\phi}_{Jl'm'}\rangle \tag{2.34}$$

are constructed, as the basis set for the Hamiltonian matrix. Here I or J denotes an atom site. For a physical basis set, the functions $\{|\chi_{Ilm}\rangle\}$ should be smooth over the whole system, while the functions $\{|\phi_{Ilm}\rangle, |\dot{\phi}_{Jl'm'}\rangle\}$ are defined independently at each atomic site. The above requirement of the global smoothness on $\{|\chi_{Ilm}\rangle\}$ determines the coefficients $h_{Ilm,Jl'm'}$ *uniquely*. This is the fundamental concept of the LMTO method. In practical calculations, $\{|\chi_{Ilm}\rangle\}$ is usually replaced by

$$|\tilde{\chi}_{Ilm}\rangle \equiv |\phi_{Ilm}\rangle + \sum_{Jl'm'} h_{Ilm,Jl'm'}\left\{|\dot{\phi}_{Jl'm'}\rangle + \tilde{o}_{Jl'm'}|\phi_{Jl'm'}\rangle\right\}. \tag{2.35}$$

The parameters $\{\tilde{o}_{Jl'm'}\}$ are properly chosen so as to construct the 'most localized' Hamiltonian. This formulation with Eq. (2.35) is usually called tight-binding LMTO method.

The tight-binding LMTO method is now widely used in condensed matter physics. One typical application is those in non-periodic systems, such as amorphous [18], because these systems can not be treated with a small simulation cell. Another typical application is the theoretical connection with the novel methods for strongly correlated systems. A recent example is the connection of the LDA method with the dynamical mean field theory (DMFT), known as 'LDA+DMFT' [19, 20].

Locality and universality of tight-binding Hamiltonian

Hereafter, in this section, we will see that the LMTO theory gives the universality of the tight-binding Hamiltonians. After a complicated procedure, we can obtain one simple resultant tight-binding Hamiltonian on an orthogonal basis set $\{|Ilm\rangle\}$. Its explicit form is given by

$$H^{(\mathrm{TB})} = C + \Delta^{1/2}\tilde{S}\Delta^{1/2} \tag{2.36}$$

$$\tilde{S} \equiv S(1 - Q^{1/2}SQ^{1/2})^{-1}. \tag{2.37}$$

Here the matrices C, Δ and Q are diagonal, for example,

$$C_{Ilm,Jl'm'} \equiv C_{Ilm}\delta_{IJ}\delta_{ll'}\delta_{mm'}, \tag{2.38}$$

while the matrices S and $\tilde{S}$ has off-diagonal elements. The matrices S and $\tilde{S}$ are called 'bare' and 'screened' structure constants, respectively. The 'bare' structure constant S is defined by

$$S_{Ilm,J'l'm'} \propto \left(\frac{1}{R_{IJ}}\right)^{l+l'+1} Y^*_{l+l',m'-m}(\hat{\boldsymbol{R}}_{IJ}), \tag{2.39}$$

where $\boldsymbol{R}_{IJ} \equiv \boldsymbol{R}_I - \boldsymbol{R}_J$ is the interatomic vector between the I-th and J-th atoms. Here, in Eq. (2.39), we drop some non-essential factors. Though we have skipped all the technical details, we can see that the *interatomic* interaction in the tight-binding Hamiltonian $H^{(\mathrm{TB})}$ is determined by the screened structure constant $\tilde{S}$. If the tight-binding Hamiltonian shows a short-range behavior, it should be reduced to the short-range behavior of $\tilde{S}$. In general, an inversed matrix A^{-1} is not sparse, even if the matrix A itself is sparse. Since, in Eq. (2.41), $\tilde{S}$ is obtained by the inversed matrix $(1 - Q^{1/2}SQ^{1/2})^{-1}$, the short-range property of $\tilde{S}$ is not trivial. The meaning of the screened structure constant can be described below; When we rewrite Eq. (2.37) as

$$\tilde{S} = S + SQ\tilde{S}, \tag{2.40}$$

Eq. (2.40) is a Dyson-like equation for $\tilde{S}$, or a self-consistent scattering problem. If orbital suffices are ignored, the explicit matrix formula is given by

$$\tilde{S}_{IJ} = S_{IJ} + \sum_K S_{IK}Q_{KK}\tilde{S}_{KJ} \tag{2.41}$$

with the atom suffices I, J, K. The second term of Eq.(2.41) is a scattering path of $J \to K \to I$. Here K denotes all the atoms in the system. The short-range property of $\tilde{S}$ is achieved by the screening effect, due to the multiple scattering.

Now the most important consequence from the LMTO theory is that the value of the screened structure constants are *universal* in the sense that they are defined for atomic structures, not for specific elements. This can be interpreted as the statement that the screening effect of the multiple scattering, described in Eq. (2.40), is essentially governed by the *geometry* of atomic structures, not by the character of each element. This universality is the origin of the 'rigid band' picture, or the tendency that the elements among the same group form similar atomic structures and electronic structures.

The short-range property of the screened structure constant $\tilde{S}$ is demonstrated in Fig. 2.1, which shows the case in the ssσ form ($\tilde{S}_{ss\sigma}$). The data are plotted as the function of the scaled interatomic distance d in units of the Wigner-Seitz radius w, among simple cubic (SC), body-centered cubic (BCC) and face-centered cubic (FCC) structures. The Wigner-Seitz radius w is defined to be the radius of the sphere whose volume equals to that per atom (v_0);

$$\frac{4\pi}{3}w^3 = v_0. \tag{2.42}$$

The solid line is an interpolation line. We can see that the screened structure constants are well localized and are well interpolated by a 'universal' curve. This means that the screening effect of the multiple scattering is well scaled by the Wigner-Seitz radius. In other words, the scattering wave 'shrinks', due to the screening effect, into a localized region that is scaled by the Wigner-Seitz radius. The shrinkage of the wave function by the multiple scattering is analogous to that by a potential wall. The universal short-range behavior of the screened structure constant directly gives the short-range tight-binding Hamiltonian, in Eq. (2.36). In short, the interatomic hoppings can be described by the universal curve in a scaled length and energy.

The shrinkage of a scattering wave is directly seen, for example, in Fig.6 of Ref. [13]; a spherical scattering wave centered on an atom shrinks within a BCC structure. The screening or shrinking behavior is formally analogous to the screening of the electrostatic field, in which a monopole field at an atom is screened due to the induced dipole field at neighbor atoms. Of course, the two situations are different, at least, in the sense that the scattering wave is described by the Helmholtz equation, Eq. (2.26), while the electrostatic field is described by the Laplace equation ($\Delta\phi = 0$).

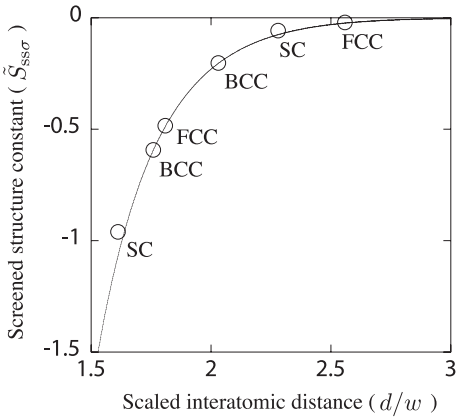


Figure 2.1: Screened structure constant in the ssσ symmetry ($\tilde{S}_{ss\sigma}$) as the function of the interatomic distance d scaled by the Wigner-Seitz radius w. The data are plotted for the first and second nearest neighbor sites among simple cubic (SC), body-centered cubic (BCC) and face-centered cubic (FCC) structures. The solid line is an interpolation line. The graph is plotted based on the data in Ref.[13].

In conclusions, the environmental effect of condensed matters is reduced to the screening effect due to the multiple scatterings, which gives the short-range tight-binding Hamiltonian from the *ab initio* theory. The screening effect is *universal*

in the following two meanings; (i) The screening effect is universal among different elements in a same structure, (ii) The screening effect is universal among different structures. This is the origin of the universality of tight-binding Hamiltonian.

In Chapter 3, we will investigate the group IV elements, systematically, within tight-binding Hamiltonian forms. The universality discussed in this section will be used as the tendency that the ratio of the interatomic hoppings $\{V_{\mathrm{ss}\sigma}, V_{\mathrm{sp}\sigma}, V_{\mathrm{pp}\sigma}, V_{\mathrm{pp}\pi}\}$ is almost unchanged among the elements in a same structure. Note that recent developments in the MTO theory will be discussed in Section 3.3.

2.3 Generalized Wannier states

This section introduces the generalized Wannier state, which was developed by Walter Kohn [21, 22] in the context of large-scale calculations. Its formulation is a generalization of the (conventional) Wannier states [23, 24].

Foundations

The generalized Wannier states $\{\phi_i\}$ is defined as localized wave functions that satisfy the equation

$$H_{\rm eff}|\phi_i\rangle = \sum_{j=1}^{\rm occ} \varepsilon_{ij}|\phi_j\rangle. \tag{2.43}$$

and the orthogonality

$$\langle\phi_i|\phi_i\rangle = \delta_{ij}. \tag{2.44}$$

Equation (2.43) has been derived in the previous section (Section 2.1) as Eq. (2.9). As explained in Section 2.1, the solutions of Eq. (2.43) is equivalent to the unitary transformation of the eigen states $\{\phi_j^{\rm (eig)}\}$

$$|\phi_i\rangle = \sum_{j}^{\rm occ.} U_{ij}|\phi_j^{\rm (eig)}\rangle, \tag{2.45}$$

where U_{ij} is a unitary matrix. The Hamiltonian matrix with respect to Wannier states

$$\langle\phi_i|H_{\rm eff}|\phi_j\rangle \tag{2.46}$$

has non-zero off-diagonal elements unlike that with eigen states. It is crucial that the generalized Wannier states reproduce the one-body density matrix, Eq. (2.16), and any physical quantity in the trace form of Eq. (2.19).

Here we derive the conventional Wannier states as a specific case of Eq. (2.45). In periodic systems, eigen states are Bloch states $\{\psi_{\nu\boldsymbol{k}}^{\rm (Bloch)}\}$ with the suffices of the band ν and the k-point $\boldsymbol{k}$, the point in the Brillouin zone. Within an isolated single band, the Wannier states can be defined $W_{\nu\boldsymbol{l}}$ with the suffices of the band ν and the lattice vector $\boldsymbol{l}$;

$$W_{\nu\boldsymbol{l}}(\boldsymbol{r}) = \int d\boldsymbol{k}\, e^{-i\boldsymbol{k}\boldsymbol{r}}\psi_{\nu\boldsymbol{k}}^{\rm (Bloch)}(\boldsymbol{r}), \tag{2.47}$$

where the integration is done within the Brillouin zone. In the present context, the corresponding unitary matrix U is given by

$$U_{ij} \Rightarrow U_{\nu\boldsymbol{l},\nu'\boldsymbol{k}} \equiv \delta_{\nu\nu'}\, e^{-i\boldsymbol{k}\boldsymbol{r}}. \tag{2.48}$$

Appendix D.1 shows a simple case with one dimensional single band case, in which the off-diagonal elements $\langle W_{\nu l}|H_{\rm eff}|W_{\nu l'}\rangle$ corresponds to the Fourier coefficients of the energy dispersion $\varepsilon_\nu(k)$. As shown in the above discussion, the original Wannier

state is given by the unitary transform only within an isolated single band ($\nu = \nu'$), while the generalized Wannier states are given by the unitary transform within different bands. ($\nu \neq \nu'$). Moreover, the concept of the generalized Wannier state can be applicable to non-periodic cases.

The concept of the generalized Wannier state is used for practical large-scale order-N methods [25, 26, 27], in which approximate Wannier states are constructed. In Section 4.2, we will derive a mean-field equation for the generalized Wannier states, which is equivalent to Eqs. (2.43) and (2.44). The mean-field equation gives practical order-N methods for constructing the generalized Wannier state within variational and/or perturbative procedures.

Discussions

Apart from large-scale calculations, the generalized Wannier state is also discussed in a different methodology. The generalized Wannier state can be constructed by the explicit unitary transforms from the eigen states, as a post-process of the conventional electronic structure calculations. This method is not an order-N method. Such methodologies are seen, from 60's, for calculation of molecules with the Hartree-Fock theory [28, 29, 30, 31]. Recently, the application with the DFT calculations was given [32], in which a measure of the delocalization is defined by the functional

$$\Omega \equiv \sum_{j}^{\text{occ.}} \left[\langle \phi_j | \boldsymbol{r}^2 | \phi_j \rangle - | \langle \phi_j | \boldsymbol{r} | \phi_j \rangle |^2 \right]. \tag{2.49}$$

The unitary transformations are done iteratively, so as to minimize the above functional. Figure 2.2 shows the results of crystalline silicon. The resultant Wannier state $|\phi_i\rangle$ is well localized as a bonding orbital. The suffix i of the Wannier state $|\phi_i\rangle$ indicates the bond site as its localization center. It is also important that the Wannier state has a node on the neighboring bond sites, because of the orthogonality to the other Wannier states whose centers are located on the neighboring bond sites.

Within simple molecules, a semi-qualitative process may be also possible, to construct the Wannier state. Such an example is given in Ref. [31], which is reviewed in Ref. [33]. Here a water molecule (H_2O) is discussed. So as to construct the Wannier states, the four valence eigen states, labeled with $(2a_1), (3a_1), (1b_2), (1b_1)$, are transformed successively only in the three steps. In each step, the unitary transformation is done between the selected two states, in a two dimensional rotation form with an empirically chosen rotational angle;

$$\text{(i)} \quad \begin{pmatrix} (l\mathrm{a}_1) \\ (b\mathrm{a}_1) \end{pmatrix} \equiv \frac{1}{5} \begin{pmatrix} 4 & 3 \\ -3 & 4 \end{pmatrix} \begin{pmatrix} (3\mathrm{a}_1) \\ (2\mathrm{a}_1) \end{pmatrix} \tag{2.50}$$

$$\text{(ii)} \quad \begin{pmatrix} (l_1) \\ (l_2) \end{pmatrix} \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} (1\mathrm{b}_1) \\ (l\mathrm{a}_1) \end{pmatrix} \tag{2.51}$$

$$\text{(iii)} \quad \begin{pmatrix} (b\mathrm{OH}_1) \\ (b\mathrm{OH}_2) \end{pmatrix} \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} (1\mathrm{b}_2) \\ (b\mathrm{a}_1) \end{pmatrix}. \tag{2.52}$$

The resultant Wannier states are labeled with $(l_1), (l_2), (b\mathrm{OH}_1)$ and $(b\mathrm{OH}_2)$. They are shown in Table 2.1, which are quite similar to those constructed by the quantitative method [31, 33]. We can assign the two lone pair orbitals $(l_1), (l_2)$ and the

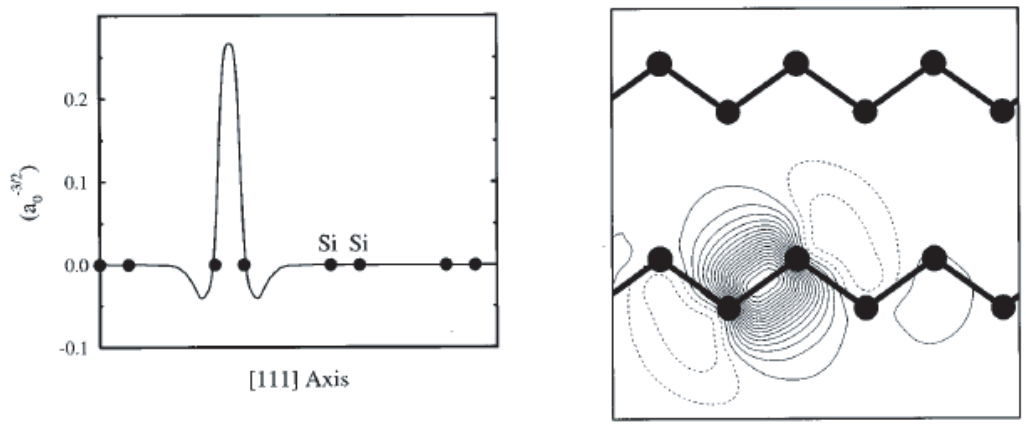


Figure 2.2: The Wannier states of bulk silicon [32] (a) Profile along the Si-Si bond. (b) Contour plot in the (110) plane of the bond chains. [N. Marzari and D. Vanderbilt, Phys. Rev. **B56**, 12847 (1997)].

two bonding orbitals $(b\mathrm{OH}_1), (b\mathrm{OH}_2)$ between the oxygen and hydrogen atoms. In Chapter 4, we will discuss that one of the large-scale order-N methods is an iterative procedure with the Wannier states (the variational order-N method). The knowledge of approximate Wannier states, such as those in Table 2.1, will be important as reliable initial wave functions in the iterative procedure.

Finally, we comment on other related theoretical concepts. There are several proposals to describe 'chemical' bonding orbitals in condensed matters. We pick out such two concepts called 'crystal orbital overlap populations' (COOP) [34] and 'crystal orbital Hamiltonian populations' (COHP) [35]. Though their mathematical formulations are different from the generalized Wannier states, such concepts should be related to the present one.

	O(2s)	O($2p_x$)	O($2p_y$)	O($2p_z$)	H1(1s)	H2(1s)
(l_1)	0.655	-0.382		0.707	-0.103	-0.103
(l_2)	0.655	-0.382		-0.707	-0.103	-0.103
$(b\mathrm{OH}_1)$	0.235	0.409	0.412		0.565	-0.162
$(b\mathrm{OH}_2)$	0.235	0.409	-0.412		-0.162	0.565

Table 2.1: Approximate Wannier states in water molecule (H_2O) [31, 33]; the four Wannier states $(l_1), (l_2), (b\mathrm{OH}_1), (b\mathrm{OH}_2)$ are expressed as the coefficients of the 2s,$2p_x$,$2p_y$,$2p_z$ orbitals of the oxygen atom and the 1s orbitals of the two hydrogen atoms. Here we ignore the small contribution by the 1s orbital of the oxygen atom.

2.4 Fundamentals of order-N methods

As discussed in Chapter 1, the order-N method is a general concept for large-scale calculations, in which the computational cost is proportional to the system size. This section describes the fundamental concept of the practical methodology.

The present purpose for large-scale calculations is restricted to structural properties, that is, the total energy calculation and the molecular dynamics. This restriction is crucial for practical order-N methods, when the total electronic structure energy $E_{\rm tot}$ does not depend explicitly on off-diagonal long-range components of the one-body density matrix ($\rho(\boldsymbol{r}, \boldsymbol{r}')$). Recently, as the foundation of the order-N method, Walter Kohn proposed a general concept called 'nearsightedness principle' [36], which is closely related to the Hohenberg-Kohn theorem [1]. Now we can recall that the Hohenberg-Kohn theorem gives the ground state energy as a functional of the charge density $n(\boldsymbol{r})$ and that the charge density is the *diagonal* elements of the density matrix ($n(\boldsymbol{r}) = \rho(\boldsymbol{r}, \boldsymbol{r})$). Instead of such a general theory, we discuss two typical examples; one is the free electron system and the other is a system with a short-range tight-binding Hamiltonian.

Example with free electrons

In the free electron system, the Hamiltonian is simply the kinetic energy part

$$H_0 \equiv -\frac{1}{2}\Delta \tag{2.53}$$

and the ground state is characterized by the Fermi wavenumber $k_{\rm F}$. The details of the following calculations will be given in Appendix D.2 and hereafter only the results are shown in this section. The total energy per volume is calculated in the reciprocal space by

$$\frac{E}{V} = \int_{k<k_{\rm F}} \frac{d\boldsymbol{k}}{(2\pi)^3}\frac{1}{2}k^2 = \frac{k_{\rm F}^5}{20\pi^2}. \tag{2.54}$$

The corresponding one-body density matrix is defined as

$$\rho(\boldsymbol{r}_1, \boldsymbol{r}_2) \equiv \int \frac{d\boldsymbol{k}}{(2\pi)^3}\frac{e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1}}{\sqrt{V}}\frac{e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_2}}{\sqrt{V}} = \int \frac{d\boldsymbol{k}}{(2\pi)^3}\frac{e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)}}{V}. \tag{2.55}$$

Due to the uniform property, the density matrix is reduced to that of the function of the distance $r \equiv |\boldsymbol{r}_1 - \boldsymbol{r}_2|$. Without the volume factor $1/V$, we redefine the density matrix and calculate

$$\begin{aligned}\rho(r) &\equiv \int_{k<k_{\rm F}} \frac{d\boldsymbol{k}}{(2\pi)^3}e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \\ &= \frac{2}{(2\pi)^2}\left\{-\frac{k_{\rm F}}{r^2}\cos k_{\rm F}r + \frac{1}{r^3}\sin k_{\rm F}r\right\}.\end{aligned} \tag{2.56}$$

The resultant density matrix shows a long-range oscillation with the Fermi wave number $k_{\rm F}$

$$\rho(r) \propto \frac{\cos k_{\rm F}r}{r^2} \quad (r \to \infty), \tag{2.57}$$

which is known as the Friedel oscillation. The short-range behavior, on the other hand, is given by the Taylor expansion as

$$\rho(r) = \frac{2}{(2\pi)^2}\left(C_0 - \frac{C_2}{2}r^2 + O(r^4)\right), \tag{2.58}$$

using the zero-th and second Taylor coefficients

$$C_0 \equiv \frac{k_{\rm F}^3}{6}, \quad C_2 \equiv \frac{k_{\rm F}^5}{15}. \tag{2.59}$$

The total energy per volume can be also calculated by the density matrix as

$$\frac{E}{V} \equiv \frac{1}{V}{\rm Tr}[\rho H_0] = \frac{1}{V}\int d\boldsymbol{r}_1 \left.\frac{-\Delta_{\boldsymbol{r}_1}}{2}\rho(\boldsymbol{r}_1,\boldsymbol{r}_2)\right|_{\boldsymbol{r}_1=\boldsymbol{r}_2} = \lim_{\varepsilon\to 0}\frac{E_{\rm sphere}(\varepsilon)}{(4\pi\varepsilon^3/3)} \tag{2.60}$$

Here $E_{\rm sphere}(\varepsilon)$ is the energy of a tiny (real-space) sphere with the radius of ε. Using Eq.(2.58) and the calculation

$$E_{\rm sphere}(\varepsilon) \equiv \int_{r<\varepsilon} d\boldsymbol{r}\frac{-\Delta_{\boldsymbol{r}}}{2}\rho(r) = \frac{C_2}{\pi}\varepsilon^3 + O(\varepsilon^5), \tag{2.61}$$

the energy per volume is given by

$$\frac{E}{V} = \lim_{\varepsilon\to 0}\frac{E_{\rm sphere}(\varepsilon)}{(4\pi\varepsilon^3/3)} = \frac{3C_2}{4\pi^2}. \tag{2.62}$$

With the definition $C_2 \equiv k_{\rm F}^5/15$, the above result reproduces that of Eq. (2.54). This shows that the total energy is determined only by the second order Taylor coefficient C_2. This statement is understandable, because the present Hamiltonian, the Laplacian operator, is the *second* order derivative. In other words, the total energy is determined explicitly by the *short-range* behavior of the density matrix. Here it should be emphasized that the above density matrix $\rho(r)$ has the off-diagonal long-range components, as in Eq. (2.57), and the system is metallic. In short, the total energy is governed by the *short-range* behavior of the density matrix, while the transport property is governed by the off-diagonal *long-range* behavior.

Example with short-range tight-binding Hamiltonians

Now we turn to the second example, a system with a short-range tight-binding Hamiltonian H. This example is directly related to the calculations in the present thesis. In such a system, the total electronic structure energy is given in the form of

$$E_{\rm elec} = {\rm Tr}[\rho H] \tag{2.63}$$

with the one-body density matrix ρ. The explicit matrix expression is

$$E_{\rm elec} = {\rm Tr}[\rho H] = \sum_{I}^{N_{\rm A}}\sum_{J}^{N_{\rm int}}\sum_{\alpha}^{\nu}\sum_{\beta}^{\nu}\rho_{J\beta I\alpha}H_{I\alpha J\beta}, \tag{2.64}$$

where (I,α) or (J,β) denotes an atomic orbital. $N_{\rm A}$ is the number of atoms in the system ($I = 1, 2..., N_{\rm A}$). The number of orbitals per atom is given by ν. For

example, $\nu = 4$ for the minimal tight-binding Hamiltonian with the s and p orbitals ($|\mathrm{s}\rangle, |\mathrm{p}_x\rangle, |\mathrm{p}_y\rangle, |\mathrm{p}_z\rangle$). Since the tight-binding Hamiltonian is short-range, the number of atoms for non-zero Hamiltonian matrix elements (N_{int}) is finite for each atom. The number of atoms (N_{int}) is, typically, that of the first or second nearest neighbor atoms. The calculation of the trace in Eq. (2.64) requires the computational cost of

$$\nu^2 N_{\mathrm{int}} N_{\mathrm{A}}. \tag{2.65}$$

Here we can say, again, that the total energy in Eq. (2.64) is determined explicitly by the *short-range* behavior of the density matrix, due to the short-range property of the Hamiltonian. For comparison, we discuss a classical model with a short-range two-body potential, in which the potential energy is given by

$$E_{\mathrm{classical}} = \sum_{I}^{N_{\mathrm{A}}} \sum_{I}^{N_{\mathrm{int}}} U_{IJ}, \tag{2.66}$$

where U_{IJ} is the two-body potential between the I-th and J-th atoms. The summation in Eq. (2.66) requires the computational cost of

$$N_{\mathrm{int}} N_{\mathrm{A}}. \tag{2.67}$$

Here one can find that the computational cost of the present electronic structure calculation should be, at least, ν^2 times larger than that of a classical model. We note that the prefactor ν^2 originates from the orbital freedom, that is, the quantum mechanical freedom. This prefactor will be discussed, in Section 5.4, as an essential point for the parallelization of the order-N method.

Discussions

The above two cases are typical examples of the situation in which the total electronic structure energy does not depend explicitly on off-diagonal long-range components of the one-body density matrix. As is seen in the beginning of the this section, the above situation is suitable for order-N methods to calculate the total energy, because the order-N method reproduces only the *short-range* behavior of the density matrix, at a sacrifice of the accuracy in the *long-range* behavior. The above discussion warns us that we should be careful, when we calculate various quantities $\langle \hat{X} \rangle$ in an order-N method. If the operator $\hat{X}$ has off-diagonal long-range components, the calculation of $\langle \hat{X} \rangle$ may be in a poor accuracy with the order-N method. Such a situation appears, when we calculated the spatial spread of a Wannier state ϕ_i

$$\langle \phi_i | (\hat{r} - \bar{r}_i)^2 | \phi_i \rangle \tag{2.68}$$

from its localization center $\bar{r}_i = \langle \phi_i | \hat{r} | \phi_i \rangle$ [27]. The calculation was done with or without an explicit localization constraint on the Wannier state ϕ_i. Since the operator $(\hat{r} - \bar{r}_i)^2$ is a long-range operator, the value of $\langle \phi_i | (\hat{r} - \bar{r}_i)^2 | \phi_i \rangle$ is more sensitive to the localization constraint than that of the energy $\langle \phi_i | H | \phi_i \rangle$.

Here the *long-range* behavior of the density matrix is discussed. With a finite temperature form, the density matrix of the free electrons, Eq. (2.56), is modified

into

$$\rho(r) \equiv \int_{k<k_{\rm F}} \frac{d\boldsymbol{k}}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} f_\tau\left(\frac{k^2}{2}\right), \tag{2.69}$$

where $f_\tau(\varepsilon)$ is the Fermi-Dirac function

$$f_\tau(\varepsilon) = \frac{1}{1+e^{(\varepsilon-\mu)/\tau}} \tag{2.70}$$

with a finite temperature τ and the chemical potential μ. An analytic evaluation [37] gives the resultant density matrix $\rho(r)$ with an additional exponential decay factor of

$$\exp\left[-\left(1+\sqrt{2}\right)\frac{\tau}{2\varepsilon_{\rm F}} k_{\rm F} r\right], \tag{2.71}$$

where $\varepsilon_{\rm F} \equiv k_{\rm F}^2/2$ is the Fermi energy. The above analysis shows that the long-range oscillation, in Eq. (2.57), at the exact ground state ($\tau = 0$) originates from the discontinuity of the occupation number at the Fermi level

$$f_\tau(\varepsilon) \to \theta(\varepsilon_{\rm F} - \varepsilon) \quad \text{in} \quad \tau \to 0. \tag{2.72}$$

In the mathematical terms, such an oscillation is known as the Gibbs phenomena, which is seen in standard textbooks of applied mathematics [38]. Such a long-range oscillation will be suppressed, when the discontinuity of the occupation number at $k = k_{\rm F}$ is faded away, as discussed above.

The discontinuity of the occupation number is equivalent to the idempotency of the ground state density matrix

$$\rho^2 = \rho. \tag{2.73}$$

From the above discussion, we can say that an essential point of the order-N methods is relaxing the exact idempotency in Eq. (2.73), which will modify the off-diagonal long-range behavior of the density matrix. The relaxation of the exact idempotency is seen, not only in Kohn's paper [36] picked out in the beginning of this section, but also in the practical order-N methods that will be explained in the next section (Section 2.5).

To end up this section, we comment on the order-N methods for *transport phenomena*. So far, we have explained the fundamentals of the order-N method for *structure* or the total energy. There exist, on the other hand, order-N methods for *transport phenomena*. An example is found in Ref.[39] with the use of the Kubo formula. Since the total energy calculation is not essential in the discussion of transport phenomena, the order-N methods for the transport phenomena should be constructed with foundations different from the present ones. See the proceedings of a recent international conference [40], as an overview of the recent developments and applications. Though we do not discuss methods for transport phenomena in this thesis, the ultimate goal for large-scale electronic structure calculations should be contain both methods for structural and transport phenomena.

2.5 Practical order-N methods

Here we review several practical order-N methods for electronic structure theory applicable to molecular dynamics. This section describes only the mathematical background of each method. See review or comparison papers [41, 42, 43, 44, 45] for details, applications, theoretical generalizations and so on. We pick out the following five typical methods; [I] the density matrix method, [II] the localized orbital method, [III] the Fermi operator expansion method, [IV] the recursion or bond-order method, and [V] the orbital-free DFT method. Except the last one, we restrict the discussion, for simplicity, to a tight-binding system that is already discussed in Section 2.4, The total electronic energy is given by

$$E_{\rm elec} = {\rm Tr}[\rho H] \tag{2.74}$$

with the one-body density matrix ρ. The correct density matrix ρ at the ground state should satisfy the idempotency

$$\rho^2 = \rho. \tag{2.75}$$

The explicit matrix form is given as

$$H_{I\alpha,J\beta} \equiv \langle I\alpha|H|J\beta\rangle \tag{2.76}$$

with the suffices of atoms (I, J) and orbitals (α, β). After the review of the five methods, we will briefly discuss a general point for practical large-scale calculations.

I : Density matrix method

The 'density matrix method'[46] is a variational method in which the explicit matrix elements of the one-body density matrix ($\rho_{I\alpha J\beta}$) are the variational freedoms. The total energy to be minimized iteratively is given as

$$E^{(I)}_{\rm elec}[\rho] = {\rm Tr}[(3\rho^2 - 2\rho^3)(H - \mu)], \tag{2.77}$$

where μ is the chemical potential that should be property chosen to reproduce the total electron number. The energy gradient

$$\frac{\partial E^{(I)}_{\rm elec}}{\partial \rho_{I\alpha,J,\beta}} \tag{2.78}$$

is used for the iterative minimization procedures. The method is closely related to a procedure on the density matrix ρ

$$\rho \Rightarrow \rho^{(\rm new)} \equiv 3\rho^2 - 2\rho^3, \tag{2.79}$$

which is called 'purification' procedure [47]. If the 'old' density matrix ρ has an approximate idempotency ($\rho^2 \approx \rho$), the new 'purified' density matrix $\rho^{(\rm new)}$ can be expected to have a better approximate idempotency in the sense that each eigen value becomes closer to zero or one. The above property can be explained, when one see the function form of $y = 3x^2 - 2x^3$, as plotted in Fig. 2.3. To see how to

reach the ground state with this energy functional, we demonstrate a case in which the density matrix is given as

$$\rho \approx \sum_i f_i |\phi_i^{(\mathrm{eig})}\rangle\langle\phi_i^{(\mathrm{eig})}|, \tag{2.80}$$

where the occupation numbers $\{f_i\}$ are nearly equal to be one or zero and the chemical potential μ is property chosen. The occupation number f_i of an eigen state contribute the energy by

$$(3f_i^2 - 2f_i^3)(\varepsilon_i - \mu). \tag{2.81}$$

For an occupied eigen level ($\varepsilon_i < \mu$), the energy function in Eq. (2.81) has a minimum at $f_i = 1$, while, for an unoccupied eigen level ($\varepsilon_i > \mu$), the energy function in Eq. (2.81) has a minimum at $f_i = 0$. Therefore, one can expect that the occupation number will be convergent to one ($f_i \to 1$) for occupied levels and zero ($f_i \to 1$) for unoccupied levels. In both cases, however, one may find that the above minimum is not the *global* minimum, but a *local* minimum. The energy function in Eq. (2.81) diverges at $f_i = \infty$ or $f_i = -\infty$, which is sometimes called 'runaway solution'. The above unphysical 'runaway' solution means a large deviation from the approximate idempotency ($\rho^2 \approx \rho$). In practical program codes, such a large deviation can be avoided by introducing an additional inner loop of the purification procedure, given by Eq. (2.79), in the iterative energy minimization procedures. This inner loop is done iteratively till the density matrix ρ recovers an approximate idempotency ($\rho^2 \approx \rho$) within a satisfactory deviation from the exact one. It should be noted that, for the convergence to the correct ground state, the initial density matrix ρ should be properly prepared, which is a common problem among iterative methods.

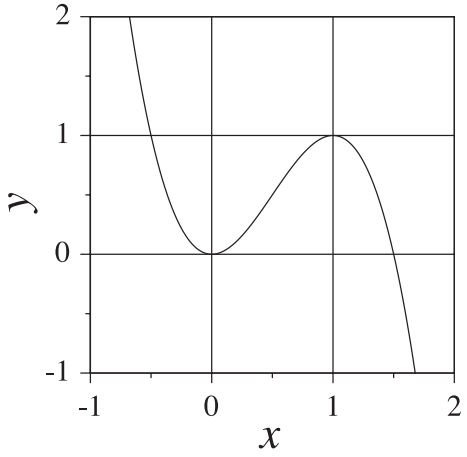


Figure 2.3: The function $y \equiv 3x^2 - 2x^3$, which is used in the density matrix order-N method.

II : Localized orbital method

The 'localized orbital method' [25, 26, 27] is directly related to the theory in this thesis. The present 'localized orbital' is the generalized Wannier state. The total

energy is given by the functional, with respect to occupied one electron states $\{\phi_i\}$, as

$$E^{(\mathrm{II})}_{\mathrm{elec}}[\{\phi_i\}] = \sum_{i,j}^{\mathrm{occ.}} A_{ji}\langle\phi_i|H-\eta_{\mathrm{s}}|\phi_j\rangle \tag{2.82}$$

$$A_{ij} \equiv 2\delta_{ij} - S_{ij} \tag{2.83}$$

where $S_{ij} \equiv \langle\phi_i|\phi_j\rangle$ is the overlap matrix between occupied states. Here the wave functions $\{\phi_i\}$ are *not* under the orthogonal constraint ($S_{ij} \neq \delta_{ij}$). The energy parameter η_{s} is what we call 'energy shift parameter'. The value of η_{s} should be chosen large sufficiently, as will be explained below. A physical quantity $\langle X\rangle$ is redefined as

$$\langle X\rangle = \sum_{i,j}^{\mathrm{occ.}} A_{ji}\langle\phi_i|X|\phi_j\rangle. \tag{2.84}$$

For example, the charge density $n(\boldsymbol{r})$ is given by

$$n(\boldsymbol{r}) = \sum_{i,j}^{\mathrm{occ.}} A_{ji}\phi^*_i(\mathrm{r})\phi_{\mathrm{j}}(\mathrm{r}). \tag{2.85}$$

It is noteworthy that, if we choose the matrix A as $A = S^{-1}$, the definition in Eq. (2.84) will be reduced to the standard one, because a set of orthogonal wave functions $\{\psi_k\}$ can be constructed as

$$|\psi_k\rangle \equiv \sum_{j}^{\mathrm{occ.}} \left(S^{-1/2}\right)_{jk} |\phi_j\rangle. \tag{2.86}$$

Using the orthogonal wave functions $\{\psi_k\}$, a physical quantity is expressed by

$$\begin{aligned}
\langle X\rangle &= \sum_{k}^{\mathrm{occ.}}\langle\psi_k|X|\psi_k\rangle \\
&= \sum_{k}^{\mathrm{occ.}}\sum_{i,j}^{\mathrm{occ.}} \left(S^{-1/2}\right)_{ki} \langle\phi_i|X|\phi_j\rangle \left(S^{-1/2}\right)_{jk} \\
&= \sum_{i,j}^{\mathrm{occ.}} \left(S^{-1}\right)_{ji} \langle\phi_i|X|\phi_j\rangle,
\end{aligned} \tag{2.87}$$

which gives Eq. (2.84) with the choice of $A = S^{-1}$. The orthogonality of $\{\psi_k\}$ ($\langle\psi_k|\psi_l\rangle = \delta_{kl}$) is directly derived from Eq. (2.87) in the choice of $X = 1$. When the inversed matrix S^{-1} is expanded as

$$\begin{aligned}
S^{-1} &= \{I-(I-S)\}^{-1} \\
&= \sum_{k=0}^{\infty}(I-S)^k \\
&\approx I + (I-S) + (I-S)^2 + \cdots
\end{aligned} \tag{2.88}$$

the sum of the first two terms gives the matrix A in Eq. (2.83).

Now we see how the above choice of the matrix A works in the present energy functional. We rewrite Eq. (2.82) as

$$E_{\rm elec}^{\rm (II)}[\{\phi_i\}] = \sum_{i,j}^{\rm occ.} A_{ji}\langle\phi_i|H|\phi_j\rangle + \eta_{\rm s}\sum_{i,j}^{\rm occ.}|\langle\phi_i|\phi_j\rangle - \delta_{ij}|^2 \tag{2.89}$$

where a constant energy term is ignored. The minimization of the second term causes a feedback force to the orthogonal relation

$$\langle\phi_i|\phi_j\rangle \to \delta_{ij} \tag{2.90}$$

with a sufficiently large value of $\eta_{\rm s}$ ($\eta_{\rm s} = \infty$). We call the second term in Eq (2.89) as 'penalty' term in the sense that, when the orthogonal relation is broken in the iterative minimization procedure, the energy will increase as a 'penalty'. The name of 'penalty' is seen in Ref.[36] in similar meanings, though the mathematical formulation is different from the present context. Moreover, one can mathematically prove [25] that, when the value of $\eta_{\rm s}$ is chosen to be larger than the highest occupied level $\varepsilon_{\rm HO}$ ($\eta_{\rm s} > \varepsilon_{\rm HO}$), the energy functional in Eq. (2.82) has the correct ground state energy $E_{\rm GS}$ as a stable point

$$E_{\rm elec}^{\rm (II)}[\{\phi_i\}] \geq E_{\rm GS}. \tag{2.91}$$

Since the wave functions will be orthogonal in the correct ground state ($S_{ij} = \delta_{ij}$), the matrix A will be reduced to $A_{ij} = \delta_{ij}$ and Eq. (2.84) will be reduced to Eq. (2.13). It is essential that the minimization of the functional in Eq. (2.82) can be done *without* any explicit orthogonalization procedure, such as the constraint scheme with the Lagrange multipliers or the Gram-Schmidt orthogonalization procedure. Unlike the other methods in this section, this method does not require a calculation of the chemical potential. The total charge $N_{\rm elec}$ may be deviated from the given value $N_{\rm elec}^{(0)}$ of the electron number

$$\begin{aligned} N_{\rm elec} &\equiv \int n(\boldsymbol{r})d\boldsymbol{r} \\ &= N_{\rm elec}^{(0)} - \sum_{i,j}^{\rm occ.}|\langle\phi_i|\phi_j\rangle - \delta_{ij}|^2 \end{aligned} \tag{2.92}$$

during the iterative minimization procedure. The above deviation, however, will reduce to zero ($N_{\rm elec} \to N_{\rm elec}^{(0)}$) in the final (ground) state, due to the property of Eq.(2.90).

In Section 4.2, we will derive a mean-field equation for the generalized Wannier state, which is equivalent to the present formulation but is derived from a different theoretical background.

III : Fermi operator expansion

The Fermi operator expansion [48, 49] is based on the Chebyshev (Tschebyscheff) polynomials expansion. Within the explanation of this method, we use the explicit 'hat' notation, say $\hat{H}$, for operators. A density matrix is formally given as

$$\hat{\rho} = f_{\rm FD}(\hat{H}) \tag{2.93}$$

where $f_{\rm FD}(\varepsilon)$ is the Fermi-Dirac distribution

$$f_{\rm FD}(\varepsilon) \equiv \frac{1}{1+e^{(\varepsilon-\mu)/\tau}} \tag{2.94}$$

with a finite temperature parameter τ and the chemical potential μ. Here the above 'temperature' parameter τ may be different from the temperature of the system, since a finite value of τ is essential for the numerical stability of the polynomial expansion. An order-N algorithm is obtained, when the operator in Eq. (2.93) is expanded by a finite set of the Chebyshev polynomials $\{T_k(x)\}$ as

$$f_{\rm FD}(\hat{H}) \approx \sum_{k=0}^{n_{\rm pl}} a_k T_k \left(\frac{\hat{H}-\varepsilon_0}{W} \right), \tag{2.95}$$

where the energy parameters ε_0 and W should be chosen so that the operator

$$\hat{x} \equiv \frac{\hat{H}-\varepsilon_0}{W}, \tag{2.96}$$

a shifted and scaled Hamiltonian, has eigen values only within the range of $-1 < x < 1$. Moreover, since the original Fermi-Dirac function is non-zero for $-\infty < \varepsilon < \infty$, the function $f_{\rm FD}(\varepsilon)$ used in Eq. (2.94) should be interpreted by a truncated one within a finite energy range ($\varepsilon_{\rm min} < \varepsilon < \varepsilon_{\rm max}$). The lower boundary energy $\varepsilon_{\rm min}$ should be chosen to be less than the lowest eigen level and the higher boundary energy $\varepsilon_{\rm max}$ should be so large that the occupation is almost zero ($f_{\rm FD}(\varepsilon_{\rm max}) \ll 1$). In result, the energy range $W \equiv \varepsilon_{\rm max} - \varepsilon_{\rm min}$ is comparable to the energy band width by the given Hamiltonian H. For the above truncated function, the coefficients ($\{a_i\}$) are determined uniquely, due to the orthogonality of the Chebyshev polynomials. The order of polynomial expansion $n_{\rm pl}$ is crucial for the computational cost. Since the $n_{\rm pl}$-th order polynomial expansion corresponds to the energy resolution of $O(W/n_{\rm pl})$, an optimal polynomial order to reproduce $f_{\rm FD}(\varepsilon)$ is estimated as

$$n_{\rm pl} \approx \frac{W}{\tau}. \tag{2.97}$$

A typical example of the Chebyshev polynomial expansion of $f_{\rm FD}(\varepsilon)$ is seen in Fig.7 in Ref. [42], which describes the diamond case with $n_{\rm pl} = 40$. With Eq. (2.95) and a complete basis set $\{|\chi_i\rangle\}$

$$1 = \sum_{i}^{\rm (basis)} |\chi_i\rangle\langle\chi_i|, \tag{2.98}$$

the total electronic energy is expressed by

$$\begin{aligned} E_{\rm elec}^{\rm (III)} &= {\rm Tr}[\hat{H} f_{\rm FD}(\hat{H})] \\ &= \sum_{i}^{\rm (basis)} \langle\chi_i|\hat{H} f_{\rm FD}(\hat{H})|\chi_i\rangle \\ &= \sum_{i}^{\rm (basis)} \sum_{k=0}^{n_{\rm pl}} a_k \langle\chi_i|\hat{H} T_k(\hat{x})|\chi_i\rangle. \end{aligned} \tag{2.99}$$

Here the vectors

$$|\chi_i^{(k)}\rangle \equiv T_k(\hat{x})|\chi_i\rangle \tag{2.100}$$

can be calculated by a recurrence relation

$$|\chi_i^{(k+1)}\rangle = 2\hat{x}|\chi_i^{(k)}\rangle - |\chi_i^{(k-1)}\rangle, \tag{2.101}$$

due to the same recurrence relation of the Chebyshev polynomials

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x). \tag{2.102}$$

IV : Recursion or bond-order method

The 'recursion' or 'bond-order' method [50] means the molecular dynamics based on the recursion method [51]. The essence of the original recursion method is as follows (See Appendix D.3 for details); For an 'input' vector $|u\rangle$, the projected Green function is given as the following continued fraction

$$\langle u|\frac{1}{\varepsilon + i0 - H}|u\rangle \approx \cfrac{1}{\varepsilon - a_1 - \cfrac{b_1^2}{\varepsilon - a_2 - \cfrac{b_2^2}{\cdots \cfrac{}{\varepsilon - a_{n_\mathrm{R}} - \cfrac{b_{n_\mathrm{R}}^2}{T_{n_\mathrm{R}}(\varepsilon)}}}}}. \tag{2.103}$$

Here we restrict the discussion to real vectors and matrices. The coefficients $\{a_n, b_n\}$ for $n = 1, 2...n_\mathrm{R}$ are real and uniquely determined for the given input vector $|u\rangle$. The integer n_R, the order of the continued fraction, should be properly chosen. The function $T_{n_\mathrm{R}}(\varepsilon)$ is a complex variable and is called 'terminator'. Several explicit function forms of $T_{n_\mathrm{R}}(\varepsilon)$ are given in textbooks. It it noteworthy, if n_R is equals to the size of the Hamiltonian matrix, the coefficients $\{a_n, b_n\}$ give, formally, a tridiagonalization of the original Hamiltonian H using a unitary matrix U;

$$U^{-1}HU = \begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & \cdots & \cdots & \cdots \end{pmatrix}. \tag{2.104}$$

When all the projected Green functions ($\langle\chi_i|G(\varepsilon)|\chi_i\rangle$) are calculated for the orthogonal complete basis set $\{|\chi_i\rangle\}$, the density of states (DOS) $D(\varepsilon)$ is given as

$$D(\varepsilon) = -\frac{1}{\pi}\sum_i^{\text{basis}} \mathrm{Im}\langle\chi_i|G(\varepsilon + i0)|\chi_i\rangle. \tag{2.105}$$

Figure 2.4 shows an example of the DOS calculated by the recursion method within the tight-binding Hamiltonian [6], in which several calculations with different values of n_R are compared [52]. The system is silicon crystal with 512 atoms in the periodic cell. The energy origin is shifted downward by 0.46 eV so that the highest occupied level in the exact diagonalization is placed at $\varepsilon = 0.0$. The lowest unoccupied level

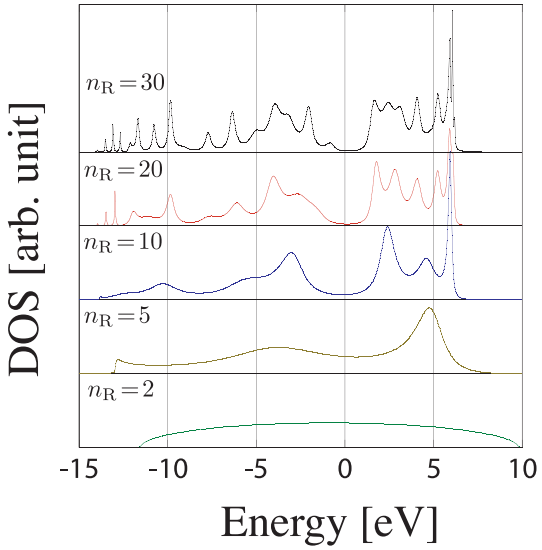


Figure 2.4: Density of states (DOS) calculated by the recursion method with several values of $n_{\rm R}$ [52]. The system is silicon crystal using a tight-binding Hamiltonian. The simulation cell contains 512 atoms. The exact diagonalization gives the highest occupied and the lowest unoccupied levels at $\varepsilon = 0.0$ and $\varepsilon = 0.8$ eV, respectively.

in the exact diagonalization is placed at $\varepsilon = 0.8$ eV. The basis set $\{\chi_i\}$ for the input vectors in the recursion method is chosen as the set of the sp^3 orbitals on all the bond sites in the diamond structure. With increasing $n_{\rm R}$, an energy gap appears in the energy region of $0 \le \varepsilon \le 0.8$ eV, as in the exact diagonalization.

For molecular dynamics, the off-diagonal elements of the density matrix (ρ_{ij}) should be calculated. From the off-diagonal matrix elements of the Green function, the density matrix is given as

$$\langle \chi_i | \rho | \chi_j \rangle = -\frac{1}{\pi} \mathrm{Im} \int d\varepsilon \, f_{\rm FD}(\varepsilon) \, \langle \chi_i | G(\varepsilon + i0) | \chi_j \rangle, \tag{2.106}$$

where $f_{\rm FD}(\varepsilon)$ is the Fermi-Dirac distribution of Eq. (2.94). The above 'temperature' parameter τ may be different from the temperature of the system, as in the Fermi operator expansion. An off-diagonal element of the Green function $\langle \chi_i | G(\varepsilon) | \chi_j \rangle$ can be calculated from the recursion method for $\langle (\chi_i + \chi_j) | G(\varepsilon) | (\chi_i + \chi_j) \rangle$, because of the relation

$$\langle \chi_i | G(\varepsilon) | \chi_j \rangle = \frac{1}{2} \left[\langle \chi_i + \chi_j | G(\varepsilon) | \chi_i + \chi_j \rangle - \langle \chi_i | G(\varepsilon) | \chi_i \rangle - \langle \chi_j | G(\varepsilon) | \chi_j \rangle \right]. \tag{2.107}$$

After the calculation of the off-diagonal elements of the Green function ($\langle \chi_i | G(\varepsilon + i0) | \chi_j \rangle$), the energy integral in Eq. (2.106) is done in a proper numerical scheme so as to obtain ρ_{ij}.

Note that the term 'bond order' is based on the fact that the basis

$$|+\rangle \equiv \frac{|\chi_i\rangle + |\chi_j\rangle}{\sqrt{2}} \tag{2.108}$$

appears in Eq. (2.107). The basis $|+\rangle$ can be interpreted as a bonding orbital, if $|\chi_i\rangle$ and $|\chi_j\rangle$ are atomic orbitals on different atom sites. In the above situation, the partial density matrix $\langle + | \rho | + \rangle$ is the occupation number of the bonding orbital $|+\rangle$.

V : Orbital-free DFT method

The 'orbital-free' DFT method is based on a foundation quite different from the other methods explained in this section. Though this method is based on the total energy functional of the DFT, given by Eq.(2.2), the kinetic energy $E_{\rm kin}$ is transformed into an explicit functional of the charge density ;

$$E_{\rm kin} \Rightarrow E_{\rm kin}[n]. \tag{2.109}$$

The resultant total energy functional $E_{\rm tot}$ is an explicit functional of the charge density $n(\boldsymbol{r})$. With the chemical potential μ, the variational equation

$$\frac{\delta}{\delta n(\boldsymbol{r})}\left\{E_{\rm tot}[n] - \mu\int n(\boldsymbol{r})d\boldsymbol{r}\right\} = 0 \tag{2.110}$$

determines the charge density $n(\boldsymbol{r})$.

A proper functional form for $E_{\rm kin}$ should be available. As a simplest form, the Thomas-Fermi function

$$E_{\rm kin}^{\rm (TF)}[n] \equiv \frac{3}{10}(3\pi^2)^{2/3}\int\{n(\boldsymbol{r})\}^{5/3}d\boldsymbol{r} \tag{2.111}$$

appears in standard textbooks or reviews, such as Refs.[53, 54, 8]. In general, a possible guiding principle for constructing the functional $E_{\rm kin}[n]$ is to reproduce the Lindhard dielectric function (See standard textbooks like Ref.[54]). The Thomas-Fermi functional reproduces the Lindhard dielectric function in the long wavelength limit. Another simple functional is the von Weizsäcker functional

$$E_{\rm kin}^{\rm (vW)}[n] = \frac{1}{8}\int\frac{|\nabla n(\boldsymbol{r})|^2}{n(\boldsymbol{r})}d\boldsymbol{r}, \tag{2.112}$$

which reproduces the Lindhard dielectric function in the short wavelength limit. From the above discussion, the sum of the two functionals

$$E_{\rm kin}[n] = E_{\rm kin}^{\rm (TF)}[n] + E_{\rm kin}^{\rm (vW)}[n] \tag{2.113}$$

seems to be a good candidate for the proper functional form. It is noteworthy that, if the second term in Eq. (2.113) is divided by nine, the functional is transformed into the simple gradient expansion form in the famous Hohenberg-Kohn paper[1]. A typical behavior of the charge density in the orbital-free DFT is shown as a figure in Ref. [55], which is reviewed in Ref. [56]. The calculated system is the neutral Kr atom and the orbital-free result is compared with the Hartree-Fock result. A characteristic feature of the orbital-free result is the lack of the quantum mechanical oscillatory structure of the radial density. See a review [56] for more discussions.

Based on the above orbital-free method, molecular dynamics simulations are performed with local pseudo potentials [57]. Since a local pseudo potential energy $E_{\rm PP}^{\rm (loc)}$, in a form of

$$E_{\rm PP}^{\rm (loc)} = \int V_{\rm PP}^{\rm (loc)}(\boldsymbol{r})n(\boldsymbol{r})d\boldsymbol{r}, \tag{2.114}$$

is an energy functional with respect to the charge density, it is applicable to the orbital-free method. The functional form for such a local potential should be constructed properly. A possible guiding principle for constructing the functional $E_{\rm PP}^{\rm (loc)}$ is, again, to reproduce the Lindhard dielectric function. A typical application is a simple metal that can be understood, basically, by the nearly free electron picture. Several improved functional forms are proposed for the kinetic energy functional and the local pseudo potential. See, for example, Ref.[58] and references therein.

A technical comment is added; In the above molecular dynamics simulations, the (valence) charge density is expanded by the plane-wave basis. Since such a program code is almost a subset of that in *ab initio* molecular dynamics with the plane-wave basis (See Appendix A.1), the programing effort may be small for the orbital-free method, if one already has a plane-wave code of *ab initio* molecular dynamics.

Discussion

So far, we have reviewed five typical order-N methods. Finally, we point out the importance of the prefactor, not the order, of the computational costs. Consider a simple case of a nearest neighbor tight-binding Hamiltonian H on a three-dimensional cubic lattice. In this system, the Hamiltonian is operated by n times successively on a vector $|\chi^{(0)}\rangle$ localized within an atom. The resultant vector

$$|\chi^{(n)}\rangle \equiv H^n|\chi^{(0)}\rangle \tag{2.115}$$

has a spatial spread only within a local region in the system, which is a desirable situation in an efficient order-N method. The above situation is true, *if the system is sufficiently large*. The resultant vector $|\chi^{(n)}\rangle$ contains $(2n+1)^3$ atoms as its spatial spread in real space. In the case with $n = 10$, for instance, the above number of atoms is $(21)^3 \approx 8000$. In a system with less than or about equal to 8000 atoms, the vector $|\chi^{(10)}\rangle$ is not localized, if without any additional approximation. On the other hand, an ideal order-N algorithm may not be required for several large-scale calculations. For example, a calculation with $O(N^2)$ computational costs can be faster than the diagonalization method with $O(N^3)$ computational costs. Therefore, the prefactor of the computational costs should be discussed for practical large-scale simulations. We have already discussed this point, in Section 2.4, in which the number of orbitals per atom contributes the prefactor of the computational cost.

Part II

Theory

Chapter 3

Tight-binding theory among elements and phases

3.1 Universal tight-binding theory

In this chapter, we investigate tight-binding formulations as a fundamental concept for simplifying the electronic structure energy. The *ab initio* theory in Chapter 2.2 gives an universal property to the tight-binding Hamiltonian in the sense that the explicit function forms of the hopping integrals are universal in the scaled energy and length unit. Such universality have been founded empirically for decades [59].

In this section, the group IV elements are focused. Among them, the diamond-structure solids are found in carbon, silicon, germanium and the low temperature solid phase of tin (α-Sn). For these materials, nearest-neighbor tight-binding Hamiltonians can be constructed within the minimal orbitals (s, p_x,p_y,p_z). Such tight-binding Hamiltonians are defined by five parameters; One is an intra-atomic parameter $\varepsilon_{\rm p}-\varepsilon_{\rm s}$, the energy difference between the atomic s and p orbitals. The other four parameters are the four interatomic hoppings ($V_{\rm ss\sigma}, V_{\rm sp\sigma}, V_{\rm sp\pi}, V_{\rm pp\pi}$,) in the Slater-Koster form [60]. The explicit numerical data in this section are calculated with a standard parametrization [61], which is systematically constructed among the group IV elements in the diamond structure solids (C, Si, Ge, α-Sn). As details, the above parametrization is carried out with the five orbitals, that is, the conventional four minimal orbitals (s, p_x, p_y, p_z) and the extra 's*' orbital [61]. The physical origin of the extra s* orbital is the spherical average of the five d orbitals. See Appendix A.2 for more explanation. The s* orbital affects mainly on the conduction band [61], while the characters of the *valence* band is not significantly changed. Moreover, the weight of the s* orbitals among occupied states is only 2.1 % in the silicon case and, at most, 2.8 % in the tin case. The above situation justifies the molecular dynamics simulation without the s* or d orbitals, since the total energy is contributed only by the valence (occupied) band. It should be noted that the essence of the present theory is based on the *universality* of the tight-binding Hamiltonian and does not depend on details of the parametrization.

	C	Si	Ge	α-Sn	β-Sn ($\leq$2.n.n.)	Pb
d [Å]	1.54	2.35	2.44	2.80	3.02 (3.17)	3.50
$N_{\rm C}$	4	4	4	4	4 (6)	12 (FCC)
$2t_{\rm sp^3}$	24.4	9.4	9.2	7.7	–	–

Table 3.1: Several data for the group IV elements; the nearest neighbor atomic distance (d), the coordination number ($N_{\rm C}$), the number of atoms within the distances of d, the hopping energy along sp^3 bonds ($t_{\rm sp^3}$). The transfer integrals are from the tight-binding formulation in Ref.[61]. In β-Sn, the values with the *second* nearest neighbor distance are also shown inside the bracket. In β-Sn, the number of atoms is six within the *second* nearest neighbor distance.

Now the tight-binding Hamiltonian is described with sp^3 orbitals. The four atomic orbitals $\{|{\rm s}\rangle, |{\rm p}_x\rangle, |{\rm p}_y\rangle, |{\rm p}_z\rangle\}$ can be transformed into the four sp^3-hybridized bases. With the above bases, the diagonal elements of the Hamiltonian matrix are given by

$$\varepsilon_{\rm h} \equiv \frac{1}{4}\varepsilon_{\rm s} + \frac{3}{4}\varepsilon_{\rm p} \tag{3.1}$$

As the off-diagonal elements, the hopping integrals between the sp^3-hybridized orbitals are classified into five types. Among them, the biggest one is that along the sp^3 bonds

$$\beta_1 \equiv \frac{1}{4}\left\{V_{ss\sigma} - 2\sqrt{3}V_{sp\sigma} - 3V_{pp\sigma}\right\}. \tag{3.2}$$

The other interatomic hoppings will be given in Section 5.5. We define the amplitude of the integral as

$$t_{sp^3} \equiv |\beta_1|. \tag{3.3}$$

If only t_{sp^3} is included as the off-diagonal elements, the resultant Hamiltonian is decomposed into each bond site as

$$\begin{pmatrix} \varepsilon_h & -t_{sp^3} \\ -t_{sp^3} & \varepsilon_h \end{pmatrix}. \tag{3.4}$$

The solution is the bonding and antibonding orbitals with the sp^3-hybridized orbitals, $|b\rangle$ and $|s\rangle$, whose energies are given by

$$\varepsilon_b \equiv \varepsilon_h - t_{sp^3} \tag{3.5}$$

$$\varepsilon_a \equiv \varepsilon_h + t_{sp^3}, \tag{3.6}$$

respectively. The resultant Hamiltonian is diagonal with respect to the above bonding and antibonding orbitals ;

$$H_0 = \sum_{i=1} \left(|b_i\rangle\varepsilon_b\langle b_i| + |a_i\rangle\varepsilon_a\langle a_i|\right), \tag{3.7}$$

where i denote the bond site in the diamond structure. The energy gap $\varepsilon_b - \varepsilon_a = 2t_{sp^3}$ is the physical origin of the covalent bondings and, therefore, the stability of the diamond structure. Figure 3.1 shows several data among the group IV elements. Here we observe that the transfer energy $2t_{sp^3}$ is a monotonically decreasing function of the nearest neighbor atomic distance d or the principal quantum number. Therefore, the diamond structure will be unstable with increasing the principal quantum number, as observed experimentally. In the tin case, the diamond structure (α phase or gray tin) appears in the low temperature phase ($T \leq 286$K), but the room temperature phase (β phase or or white tin) has a different structure called 'β-Sn structure'. In the lead case, the next heavier element in the group IV elements, the FCC structure appears with metallic properties.

In the present context, one can think of the β-Sn structure as an intermediate case between the diamond structure and the FCC structures. The nearest neighbor atomic distance in β-Sn (3.02 Å) is larger than that in α-Sn (2.80 Å) but smaller than that in Pb (3.50 Å). Moreover, in the β-Sn structure, each tin has two *second* nearest neighbor atoms that are slightly further away (3.17 Å) than the first nearest neighbor atoms, as in Table 3.1. If we ignore the above difference of the first and second nearest neighbor distances, the coordination number N_C will be defined as six.

One-parameter theory among elements

The tight-binding theory results in the fact that the ratio among these four interatomic hoppings ($V_{\mathrm{ss}\sigma}, V_{\mathrm{sp}\sigma}, V_{\mathrm{sp}\pi}, V_{\mathrm{pp}\pi},$) is almost unchanged among the elements. With this universality, the number of the independent energy parameters is reduced to only *two*. Now we can pick out $(\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}})$ and $(\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}})$ as the two independent parameters. These are equivalent to β_0 and $t_{\mathrm{sp^3}} = |\beta_1|$, respectively. Now a ratio

$$\alpha_{\mathrm{m}} \equiv \frac{\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}}}{\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}} \equiv \frac{\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}}}{2t_{\mathrm{sp^3}}}. \tag{3.8}$$

is defined as the unique parameter that explains the chemical trend among the group IV elements. Such a one-parameter theory can be found in several reviews and textbooks [59, 62]. The parameter α_{m} in Eq. (3.8) is called 'metallicity'. The value of α_{m} increases monotonically from lighter atoms into heavier atoms. The explicit values in the present parameterizations [61] are $\alpha_{\mathrm{m}} = 0.34, 0.63, 0.81, 0.91$ for C, Si, Ge, α-Sn, respectively. When the real materials are classified by the parameter α_{m}, the materials are reduced to a universal model with the energy unit scaled by the transfer energy $t_{\mathrm{sp^3}}$. Since the transfer energy $t_{\mathrm{sp^3}}$ is a monotonic function of the lattice constant (See Table 3.1), the scaling by the transfer energy $t_{\mathrm{sp^3}}$ is equivalent to the scaling by the lattice constant. Therefore, the present one-parameter description corresponds to the description with a scaled energy or length.

The above chemical trend among the group IV elements influences the character of the wave functions as the deviation from the ideal sp^3 hybridization. So as to monitor the above deviations, we define the weight of s orbitals $f_{\mathrm{s}}^{(i)}$, for each wave function ϕ_i, as

$$f_{\mathrm{s}}^{(i)} \equiv \sum_I |\langle \phi_i | I\mathrm{s} \rangle|^2, \tag{3.9}$$

where $|I\mathrm{s}\rangle$ is the s orbital on the I-th atom. Especially, f_{s} denotes the average of the occupied wave functions, which is uniquely determined for a system as

$$f_{\mathrm{s}} \equiv \sum_I \langle I\mathrm{s}|\rho|I\mathrm{s}\rangle \tag{3.10}$$

from the density matrix. Figure 3.1 shows the resultant weight of s orbitals, in which the weight of s orbitals f_{s} is a monotonically increasing function of α_{m}. The above trend is explained, as follows; In a system with a small metallicity ($\alpha_{\mathrm{m}} \ll 1$), the Hamiltonian will be quite similar to H_0 in Eq. (3.7). In such a system, the Wannier state will be a bonding orbital with the ideal sp^3 hybridization ($|\phi_i\rangle \approx |\mathrm{b}_i\rangle$). For an ideally sp^3-hybridized wave function, as $|\mathrm{b}_i\rangle$ in Eq. (3.7), the value of $f_{\mathrm{s}}^{(i)}$ is $f_{\mathrm{s}}^{(i)} = 1/4$ from its definition. Here we discuss the opposite limiting case, the case of $\alpha_{\mathrm{m}} \to \infty$. This limiting case corresponds to the dilute case, in which the interatomic hopping β_1 will be much smaller than the value of $(\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}})$ in Eq. (3.8). The resultant electronic configuration is $\mathrm{s}^2\mathrm{p}^2$, like an isolated atom, in the sense that two electrons per atom occupy the s orbital and the other two electrons occupy the p orbital. The resultant value of f_{s} is $f_{\mathrm{s}} = 1/2$. The corresponding electronic structure is a p-band metal, because the s band and the p band are well separated energetically and the s band is fully occupied. In short, the above trends among the

group IV elements can be interpreted as the change from an sp^3-hybridized insulator, in the lighter elements, into a p-band metal, in the heavier elements. This change is often called 'dehybridization', because the sp^3 hybridization is canceled and the electronic configuration is reduced to that in an isolated atom ($f_s = 1/4 \Rightarrow 1/2$).

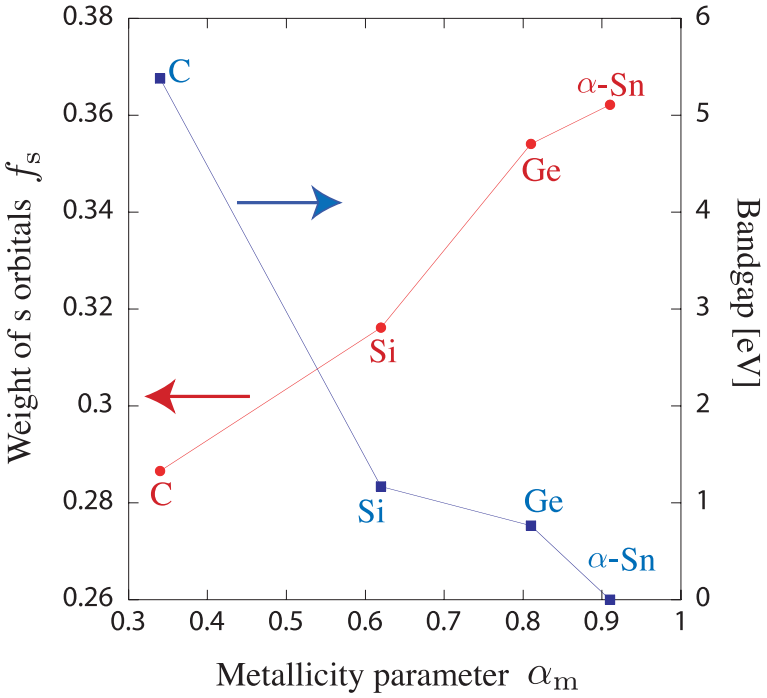


Figure 3.1: The weight of s orbitals f_s and the band gap are plotted as the function of the metallicity parameter α_m among the group IV elements. The calculations are done using a set of tight-binding Hamiltonians [61].

The band gap is estimated using the metallicity parameter α_m. The difference of the antibonding and bonding levels ($\varepsilon_a - \varepsilon_b$) corresponds to the energy difference between the band centers of the valence and conduction bands. On the other hand, the energy ($\varepsilon_p - \varepsilon_s$) motivates the transfer between bond sites and results in the finite band widths of the valence and conduction bands. Therefore, the band gap can be estimated as

$$\Delta_{\rm est} \equiv (\varepsilon_a - \varepsilon_b) - (\varepsilon_p - \varepsilon_s) = 2\, t_{sp^3}\, (1 - \alpha_m). \tag{3.11}$$

From Eq. (3.11), a system would be metallic, when $\alpha_m \to 1$. This is why the parameter α_m is called 'metallicity'. In Fig. 3.1, the band gap decreases with the increase of the metallicity, as expected from Eq. (3.11). Particularly, the metallicity in α-Sn is almost one ($\alpha_m = 0.91$) and the band gap is reduced to be zero. Since the vanishment of the band gap means the instability of the sp^3 bonding, the above result is consistent to the fact that the diamond structure is not stable in the heavier elements.

3.2 Liquid and surface phases of silicon

In this section, we will discuss that the universal tight-binding theory explains several structure among non-crystalline phases. This is directly related to the fact that practical transferable tight-binding Hamiltonians are applicable to non-crystalline phases. Here, we pick out the liquid phase and the (001) surface of silicon.

Liquid silicon

Liquid silicon is metallic with the coordination number of $N_c = 6.4$, which is well reproduced by *ab initio* [63] and tight-binding [6] molecular dynamics simulations ($N_c \approx 6.5$). We calculate the weight of s orbitals f_s in liquid silicon within the tight-binding Hamiltonian and obtain the result of $f_s = 0.43$. The resultant value is an intermediate value between that in the solid phase ($f_s = 0.36$) and that in an isolated atom ($f_s = 1/2$). Table 3.2 shows the weight of s orbitals among different elements or phases. In the liquid phase, the average of the transfer energy should be smaller than that in the solid phase, due to the disorder. From the above discussion, the melting of silicon can be explained by 'dehybridization', in which an sp^3-hybridized insulator is changed into a p-band metal.

Si(001) surface and the comparison with C and Ge (001) surfaces

Now we turn to discuss the surface reconstruction. Though carbon, silicon and germanium form the same (diamond) structure in the bulk phase, they may form different structures on reconstructed surfaces. One example is seen on the dimer geometry on the (001) surface. On the silicon case, a pair of surface atoms forms an asymmetric dimer, which is observed by tight-binding calculations [65], *ab initio* calculations [66], and experiments [67]. Figure 3.2(a) shows the geometry of the asymmetric surface dimer. Here the surface atom near the vacuum region is called 'up' atom, while the other surface atom is called 'down' atom. Figure 3.3 shows *ab initio* calculations for the carbon, silicon and germanium cases [70]. In results, a symmetric dimer appears in the carbon case, while quite similar asymmetric dimers appear in the silicon and germanium case.

The above chemical trend on the surface dimer geometry can be explained by the dehybridization mechanism. A systematic investigation is carried out among the elements with the metallicity parameter α_m. The metallicity parameter for silicon is given as $\alpha_m = 0.78$ in the present Hamiltonian [6]. Among other minimal tight-binding Hamiltonians, the values of α_m are $\alpha_m = 0.35$ for C [64], $\alpha_m = 0.75$ [68] for Si, and $\alpha_m = 0.77$, for Ge [69]. Note that the metallicity parameters of

	ideal sp^3	C	Si	l-Si	isolated atom
f_s	1/4	0.30	0.36	0.43	1/2
N_c	4	4	4	6.4	—

Table 3.2: The weight of s orbitals f_s and the coordination number N_c among crystalline carbon, crystalline silicon, and liquid silicon. The calculations are done using minimal tight-binding Hamiltonians for carbon [64] and silicon [6].

silicon and germanium are indistinguishable within the above minimal tight-binding Hamiltonians.

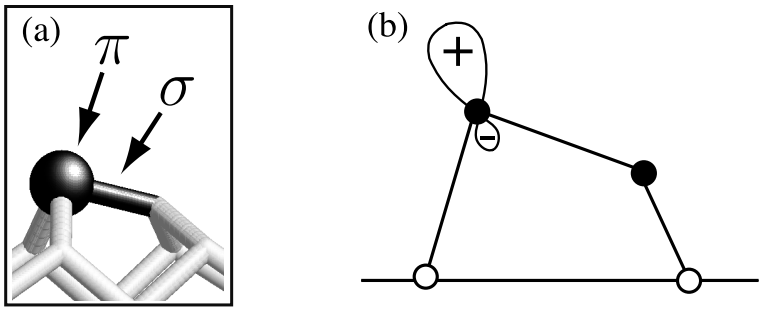


Figure 3.2: (a) A three-dimensional view of the calculated asymmetric dimer on the Si(001) surface. The bonding 'σ' state is drawn as a black rod, while the atomic 'π' state at the 'up' atom is drawn as a black ball. (b) The schematic picture of the atomic 'π' orbital localized on the 'up' atom. The surface dimer atoms are drawn as filled circle, which lie in the plane of the paper. The subsurface atoms are drawn as open circle, which do not lie in the present plane.

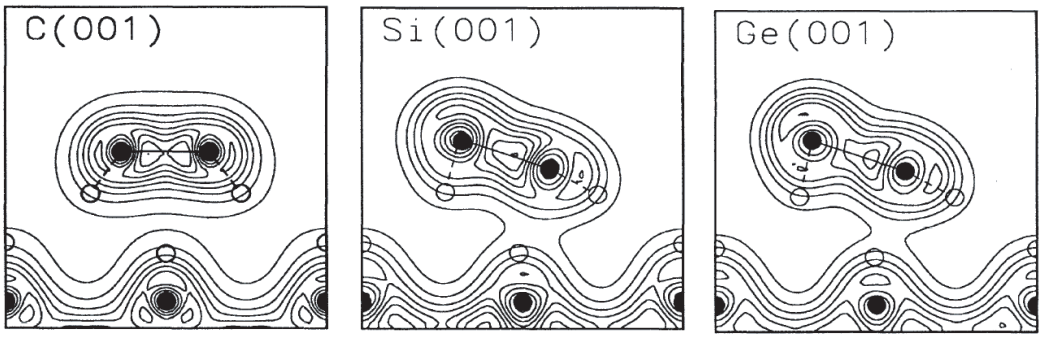


Figure 3.3: The dimer geometry of the (001) surface by *ab initio* calculations [70] [P. Kröger and J. Pollmann, Phys. Rev. Lett. **74**, 1155 (1995)]. The valence charge density is plotted within the (110) plane, in which the dimer bond lies. A black circle denotes an atom that lies within the (110) plane. A open circle denotes the projected position of an atom that does *not* lie in the (110) plane.

Before the discussion of the reconstructed surface, we explain the *unreconstructed* (001) surface. If the system is assumed to be in an ideal sp^3 bonding system, a surface atom has two dangling bond states in the ideal sp^3 hybridization. We denote the two states as $|h_1\rangle, |h_2\rangle$, which are defined as

$$|h_1\rangle \equiv \frac{1}{2}\{|s\rangle + |p_x\rangle + |p_y\rangle + |p_z\rangle\} \tag{3.12}$$

$$|h_2\rangle \equiv \frac{1}{2}\{|s\rangle - |p_x\rangle - |p_y\rangle + |p_z\rangle\}\,. \tag{3.13}$$

The other two sp^3 orbitals are parts of two 'back bonds' between the surface atoms and the subsurface atoms. From the two dangling bond states ($|h_1\rangle, |h_2\rangle$), two atomic states $|\alpha\rangle, |\beta\rangle$ can be defined as

$$|\alpha\rangle \equiv \frac{|h_1\rangle + |h_2\rangle}{\sqrt{2}} = \frac{|\mathrm{s}\rangle + |\mathrm{p}_z\rangle}{\sqrt{2}} \tag{3.14}$$

$$|\beta\rangle \equiv \frac{|h_1\rangle - |h_2\rangle}{\sqrt{2}} = \frac{|\mathrm{p}_x\rangle + \mathrm{p}_y\rangle}{\sqrt{2}}. \tag{3.15}$$

Figure 3.4 shows the schematic picture of the above two orbitals. The energies of the orbitals are given as

$$\langle\alpha|H|\alpha\rangle = \frac{\varepsilon_\mathrm{s} + \varepsilon_\mathrm{p}}{2} \tag{3.16}$$

$$\langle\beta|H|\beta\rangle = \varepsilon_\mathrm{p}. \tag{3.17}$$

Among the mixing states of the two orbitals $\{|\mathrm{h}_1\rangle, |\mathrm{h}_2\rangle\}$, the state $|\alpha\rangle$ has the lowest energy, $(\varepsilon_\mathrm{s} + \varepsilon_\mathrm{p})/2$, due to the maximum weight of s orbital ($f_\mathrm{s} = 1/2$). On the other hand, the state $|\beta\rangle$ has the highest energy, ε_p, due to the minimum weight of s orbital ($f_\mathrm{s} = 0$). In results, the two electrons in the (sp^3) dangling bond states ($|\mathrm{h}_1\rangle, |\mathrm{h}_2\rangle$) should form the lone pair state of $|\alpha\rangle$ with an energy gain of

$$\langle\alpha|H|\alpha\rangle - \varepsilon_\mathrm{h} = \frac{\varepsilon_\mathrm{s} + \varepsilon_\mathrm{p}}{2} - \frac{\varepsilon_\mathrm{s} + 3\varepsilon_\mathrm{p}}{4} = -\frac{\varepsilon_\mathrm{p} - \varepsilon_\mathrm{s}}{4}. \tag{3.18}$$

In other words, the *unreconstructed* surface is stabilized by the dehybridization mechanism ($f_\mathrm{s} = 1/4 \rightarrow 1/2$). We will observe this dehybridization mechanism, directly, as the elementary fracture process in Section 7.2.

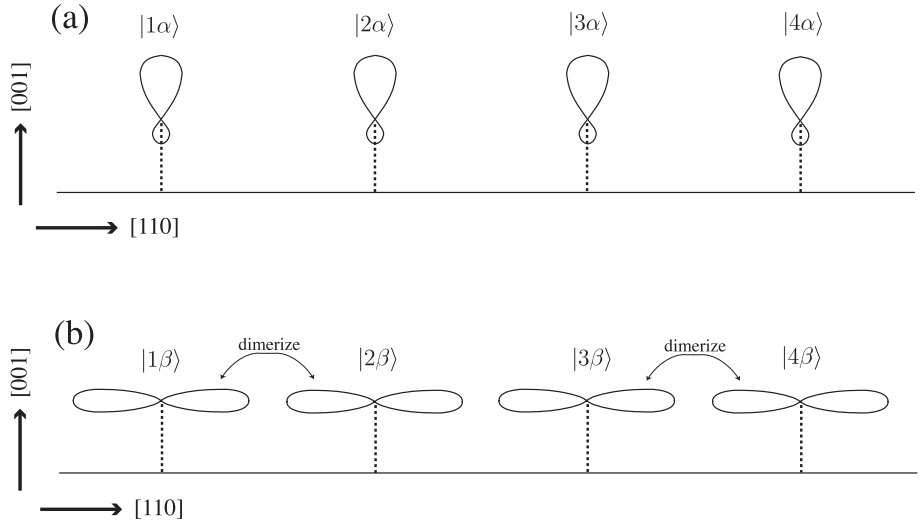


Figure 3.4: Schematic pictures of the surface states on the (001) surface of the diamond structure; (a) $|i\alpha\rangle$ and (b) $|i\beta\rangle$, where $i = 1, 2, 3, 4$ indicates the atom site, respectively. The dimerization is also shown schematically.

The surface dimerization originates mainly from the hopping between 'β' orbitals of surface atoms, as in Fig.3.4(b). In the Slater-Koster form, the interaction is

written in

$$\langle i\beta|H|j\beta\rangle = V_{\mathrm{pp}\sigma}, \tag{3.19}$$

which can be interpreted as a σ bonding. For the reconstructed surface dimer, a three-dimensional figure is given in Fig.3.2(a), in which localized states can be defined as generalized Wannier states. One localized state is the 'σ' bonding state that connects the asymmetric dimer. The origin of the bonding is mainly the interaction in Eq. (3.19). Another state is localized on the 'up' atom, the dimerized atom near the vacuum region. This localized state is sometimes called 'π' state, because the direction of its p components is nearly perpendicular to the dimer bond. The schematic picture of the atomic 'π' orbital is shown in Fig.3.2(b). The 'π' state has a larger occupation on the vacuum region, like the 'α' state in Fig.3.4. The corresponding *ab initio* wave function is seen, for example, in Fig.8(c) of Ref.[71], as one of the occupied wave function of the surface ground state. Note that, for the above 'σ' and 'π' states, the unoccupied states, 'σ^*' and 'π^* states, can be also defined. The 'σ^*' state is an antibonding state, while the 'π^* state is an atomic state localized on the 'down' atom.

We found that such (a)symmetric dimers on the (001) surface can be directly shown by the universal tight-binding theory. Figure 3.5 shows the calculated geometry of the (001) surface dimer, in which the tight-binding parameters are tuned within its universality. The actual calculations are based on the silicon tight-binding Hamiltonian, but we tune the atomic level difference ($\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}}$) so as to control the metallicity parameter α_{m}. As details, the calculations are done with alternately buckled asymmetric dimers as the initial structure, which will be seen in Fig.6.2(a) (See Section 6.3). The result shows a symmetric dimer in the carbon case ($\alpha_{\mathrm{m}} \approx 0.35$) and shows an asymmetric dimer in the silicon or germanium case ($\alpha_{\mathrm{m}} \approx 0.75 - 0.78$). In the latter case, the dimer bond is perpendicular to the plane that includes the two back bonds on the 'up' atom ($\phi \approx 90°$). These results reproduce satisfactory the *ab initio* results in Fig. 3.3.

For the energetic discussion, an energy quantity is defined as

$$\Delta\varepsilon_i^{(\mathrm{cov})} \equiv \langle\phi_i|H|\phi_i\rangle - \left[f_{\mathrm{s}}^{(i)}\varepsilon_{\mathrm{s}} + (1 - f_{\mathrm{s}}^{(i)})\varepsilon_{\mathrm{p}}\right]. \tag{3.20}$$

A negative value of $\Delta\varepsilon_i^{(\mathrm{cov})}$ corresponds to the energy gain of a covalent bonding. The 'σ' state has the gain of $\Delta\varepsilon_i^{(\mathrm{cov})} \approx -2$eV, which mainly contributes to the dimerization energy (about -2eV) [66]. The 'π' state has much smaller $\Delta\varepsilon_i^{(\mathrm{cov})}$, which is comparable to the energy difference between the asymmetric and symmetric dimers (the order of 0.1eV) [66]. As shown above, the lowest energy atomic state is obtained by the 'α' state, due to the dehybridization mechanism and the orthogonality from the back bond states. This mechanism is essential to the geometrical feature of $\phi \approx 90°$. In other words, the above geometrical feature is contributed mainly by the dehybridization mechanism of the 'π' state.

On the other hand, the symmetric dimer, as in the carbon case, is characterized by a π bond, instead of the non-bonding atomic state ($|\alpha\rangle$). This can be understood as the fact that the energy gain of the dehybridization mechanism is relatively small in carbon, due to the smallness of α_{m}. These features can be explained by the energy competition between the π bonding state that is stabilized by the transfer energy

and the partially ionic state that is stabilized by the dehybridization mechanism. In short, the dimer geometry, symmetric or asymmetric, is determined by the energetic competition between the transfer energy and the energy gain of the dehybridization mechanism, that is, to gain the energy with increasing the weight of s orbitals. Since the above energy competition is the property of the Hamiltonian matrix elements, it may be inherent in all the structures among the group IV elements. Therefore, the above chemical trend in the (001) surface dimer can be interpreted as one example of the general picture in which a double bond, σ and π bonds, is easily formed in the carbon case but is not in the silicon nor germanium cases.

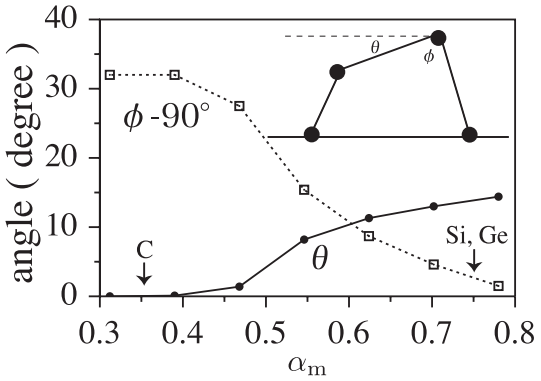


Figure 3.5: The dimer geometry of the (001) surface within the tight-binding Hamiltonians. The tilt angle θ and the angle ϕ are plotted as the function of the metallicity parameter $\alpha_{\rm m}$. The angle ϕ is defined as the angle between the surface dimer and the plane of the two back bonds of the 'up' atom. We tune the value of $(\varepsilon_{\rm p} - \varepsilon_{\rm s})$, so as to change the metallicity $\alpha_{\rm m}$ continuously.

3.3 Summary and discussion

Summary

In this chapter, several structures among the group IV elements were systematically investigated using the universality of the tight-binding Hamiltonian. The universality of the tight-binding Hamiltonian was introduced, in Section 2.2, from the *ab initio* theory. Among the elements, the universality is reduced to the fact that the tight-binding Hamiltonians for the above materials can be described by a one-parameter scaling theory, with the metallicity parameter $\alpha_{\rm m}$. The one-parameter scaling theory can be interpreted as the energy competition between the bonding mechanism and the dehybridization mechanism. The former mechanism is governed by the gain of the transfer energy, while the latter mechanism is governed by the energy gain due to the increase of the weight of s orbitals. For quantitative discussions, we monitored the weight of s orbitals ($f_{\rm s}$) for the occupied wave functions. The following points were discussed;

(I) Within the diamond structure solids (C, Si, Ge, α-Sn), the above one-parameter scaling theory is reduced to the deviation from the ideal sp^3 hybridization. As results, the sp^3-hybridized bond is stable in a lighter element but will be unstable in a heavier element, due to the dehybridization mechanism.

(II) The melting of silicon is understood by the dehybridization mechanism.

(III) The (a)symmetric dimer geometry on the (001) surface among C, Si and Ge is understood by the dehybridization mechanism. Unlike the cases (I) and (II), this dehybridization mechanism means a local mechanism within the surface region.

Among these results, we can find that the dehybridization mechanism plays the essential role in the quantum mechanical freedom of the electronic systems. Note that we will see, later in this thesis, that the dehybridization mechanism is essential also in the fracture process of silicon crystal.

Discussion

Hereafter, the topics for future works are discussed. Since the present theory is limited to properties among the group IV elements, a generalization of the present work may be a systematic investigation among the compounds in the $A^{(n)}B^{(8-n)}$ type. Such approaches can be found in many references based on empirical methods. For example, a phase diagram among this type of compounds is plotted in Ref.[72], which is reviewed in Fig.8 of Ref.[59]. In the figure, the materials are systematically classified into the two groups, the group in the fourfold coordination and that in the sixfold coordination. The classification is carried out by two parameters, that is, (i) the average of the principal quantum numbers of the elements ($\bar{n} \equiv (n_{\rm A}+n_{\rm B})/2$), and (ii) the difference of the electronegativity between the elements ($\Delta X \equiv X_{\rm A} - X_{\rm B}$). The above *two*-parameter theory corresponds to a direct theoretical extension of the present *one*-parameter theory among the group IV elements.

The generalization of the theory can be discussed also from the viewpoint of the energy functional form. The discussions in this chapter are based on the minimal

tight-binding Hamiltonian form within s and p orbitals. There should be problems that can *not* be solved within the present Hamiltonian form. The problems can be clearly seen, when the present total energy functional is compared with that in the *ab initio* one. Now we discuss such problems and several practical solutions with applications to the Si(001) surface. One problem of the minimal tight-binding Hamiltonian is the lack of d orbitals as bases. A practical solution for this problem is the use of the extra 's*' orbital [61], the spherical average of five d orbitals, as explained in Section 3.1. As an application [73], the optical spectra of the Si(001) surface was calculated and compared with experimental results. Another problem is the lack of the electrostatic energy term. Since the asymmetric dimer of the Si(001) surface is a partially ionic system, the electrostatic energy is not negligible, though it was estimated to be not essential [65] for the conclusion of the asymmetric dimer geometry. A direct correction is given by the on-site Coulomb term in the form of

$$E_{\rm Coul} \equiv U \sum_{I}^{\rm atom} \left(n_I - n_I^{(0)} \right), \tag{3.21}$$

where n_I is the calculated valence electron number of the I-th atom and $n_I^{(0)}$ is that in the neutral atom. For example, $n_I^{(0)} = 4$ in silicon. U is a given energy parameter. An application to the Si(001) surface seen in Ref.[74] with $U = 3$ eV.

Since these generalizations do *not* cause any fundamental problem in the numerical algorithm, even with the order-N methods, their applications will be possible future works. Such generalizations, however, increase the computational costs. In general, it is crucial for large-scale calculations to find a *simplest* Hamiltonian or energy functional, so as to reproduce a correct structure. Therefore, we should continue to develop the theory for constructing simple and practical Hamiltonians.

Apart from the discussion of the computational costs, a fundamental theoretical improvement should be a self-consistent construction of the tight-binding Hamiltonian during the molecular dynamics simulation. In this context, interesting approaches can be found in the new muffin-tin orbital (MTO) formulations that includes the total energy calculations. Among the MTO formulations, the one explained in Section 2.2 is usually called the second-generation MTO formulation. The new formulations are called the third-generation MTO formulations and one of them is called 'NMTO' method [75, 76, 77, 78]. Here the 'NMTO' method gives the MTO's with the N-th order in the energy, which is a generalization of the linear (LMTO) method ($N = 1$). Based on the new methodology, a bonding muffin-tin orbital was constructed in silicon crystal and the resultant wave function, Fig. 8 in Ref. [78], is quite similar to the Wannier state in Fig. 2.2. They called the wave function 'Wannier-like' MTO. As another example, a tight-binding Hamiltonian for carbon was constructed [79].

Chapter 4

Theories for large-scale calculations

4.1 Quantum mechanics with one-body density matrix

In this section, the quantum mechanical framework is reformulated with the density matrix, as a foundation of the large-scale electronic structure theory.

Density matrix

As explained in Section 2.4, a common procedure for large-scale calculations is the construction of the one-body density matrix without the matrix diagonalization. The one-body density matrix is defined as

$$\rho \equiv \sum_i f_i |\phi_i^{(\mathrm{eig})}\rangle\langle\phi_i^{(\mathrm{eig})}| \tag{4.1}$$

with eigen states $\{\phi_i^{(\mathrm{eig})}\}$ and occupation numbers $\{f_i\}$ $(0 \leq f_i \leq 1)$. The density matrix should satisfy the commutation relation

$$0 = H\rho - \rho H \tag{4.2}$$

and any physical quantity $\langle X\rangle$ is expressed as

$$\langle X\rangle = \sum_i f_i \langle\phi_i|X|\phi_i\rangle = \mathrm{Tr}[\rho X]. \tag{4.3}$$

The number of electronic states is defined by

$$N \equiv \mathrm{Tr}[\rho]. \tag{4.4}$$

Hereafter we consider the subspace division of the occupied Hilbert space. The division is done by decomposing the density matrix ρ into two subspaces of ρ_{A} and $(\rho - \rho_{\mathrm{A}})$. The division should be done within the orthogonality

$$\rho_{\mathrm{A}}(\rho - \rho_{\mathrm{A}}) = 0. \tag{4.5}$$

We will derive a general equation for a subspace ρ_{A} within several conditions. Before the construction of the general equation, we discuss the two cases of the subspace division; the subspace division with eigen states or with Wannier states.

Case (1) : Subspace division with eigen states

When the eigen states $\{\phi_i^{(\mathrm{eig})}\}$ are classified into two groups A and B, the density matrix is decomposed into the corresponding two parts

$$\rho \equiv \rho_{\mathrm{A}} + \rho_{\mathrm{B}} \tag{4.6}$$

where

$$\rho_{\mathrm{A}} \equiv \sum_i^{\mathrm{A}} f_i |\phi_i^{(\mathrm{eig})}\rangle\langle\phi_i^{(\mathrm{eig})}|, \tag{4.7}$$

$$\rho_{\mathrm{B}} \equiv \sum_i^{\mathrm{B}} f_i |\phi_i^{(\mathrm{eig})}\rangle\langle\phi_i^{(\mathrm{eig})}| = \rho - \rho_{\mathrm{A}}. \tag{4.8}$$

Here we call ρ_A and ρ_B 'subsystems'. The number of states in the two subsystems are given, respectively, by

$$N_A \equiv \mathrm{Tr}[\rho_\mathrm{A}], \tag{4.9}$$
$$N_B \equiv \mathrm{Tr}[\rho_\mathrm{B}]. \tag{4.10}$$

The commutation relations

$$0 = H\rho_\mathrm{A} - \rho_\mathrm{A}H, \tag{4.11}$$
$$0 = H\rho_\mathrm{B} - \rho_\mathrm{B}H. \tag{4.12}$$

are satisfied for the subsystems. Any physical quantity $\langle X\rangle$ is decomposed into

$$\langle X\rangle = \mathrm{Tr}[\rho_\mathrm{A}X] + \mathrm{Tr}[\rho_\mathrm{B}X]. \tag{4.13}$$

The two subsystems should be under the orthogonal constraint

$$\rho_\mathrm{A}\rho_\mathrm{B} = 0, \tag{4.14}$$

which is equivalent to the orthogonality relation

$$\langle\phi_i|\phi_j\rangle = \delta_{ij}. \tag{4.15}$$

Here we define a Hamiltonian

$$H^{(\mathrm{A})}_{\mathrm{map,I}} \equiv H + 2\eta_\mathrm{s}\rho_\mathrm{B}, \tag{4.16}$$

with an energy parameter η_s. This Hamiltonian satisfies the commutation relation

$$0 = H^{(\mathrm{A})}_{\mathrm{map,I}}\rho_\mathrm{A} - \rho_\mathrm{A}H^{(\mathrm{A})}_{\mathrm{map,I}}, \tag{4.17}$$

due to Eqs. (4.11) and (4.14). We call the Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map,I}}$ the mapped 'type I (one)' Hamiltonian. We also call the parameter η_s 'energy-shift parameter'. Equations (4.17) and (4.9) are in the same form as in Eqs. (4.2) and (4.4). If the 'B' subsystem (ρ_B) is given, the problem for the 'A' subsystem (ρ_A) is mapped to a standard quantum mechanical problem with the well defined Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map,I}}$ and the electron number N_A.

With a sufficiently large value of $\eta_\mathrm{s}(\eta_\mathrm{s}\to\infty)$, the 'A' subsystem is the ground state of the following energy functional

$$\begin{aligned} E^{(\mathrm{A})}_{\mathrm{map,I}}[\rho_A] &\equiv \mathrm{Tr}[H^{(\mathrm{A})}_{\mathrm{map,I}}\rho_\mathrm{A}] \\ &= \mathrm{Tr}[H\rho_\mathrm{A}] + 2\eta_\mathrm{s}\mathrm{Tr}[\rho_\mathrm{A}\rho_\mathrm{B}] \\ &= \mathrm{Tr}[H\rho_\mathrm{A}] + 2\eta_\mathrm{s}\sum_i^{\mathrm{A}}\sum_j^{\mathrm{B}} f_i f_j |\langle\phi_i|\phi_j\rangle|^2. \end{aligned} \tag{4.18}$$

The second term will be zero at its minimum, when the orthogonality ($\rho_\mathrm{A}\rho_\mathrm{B} = 0$) is satisfied. The second term is similar to the 'penalty functional' in Ref. [36]. If the orthogonality is not satisfied, this energy term gives an increasing positive value as a 'penalty'. Therefore, the minimization procedure of $E^{(\mathrm{A})}_{\mathrm{map,I}}$ with respect to ρ_A is mapped to the minimization within the subspace that is orthogonal to the subspace of ρ_B. This is the fundamental concept of dividing the occupied Hilbert space.

Case (2) : Subspace division with Wannier states

Now we turn to the case in which the subsystems are concerned with Wannier states, not eigen states. As explained in Section 2.3, Wannier states are defined by the unitary transformation of the occupied eigen states;

$$|\phi_i^{(\mathrm{WS})}\rangle \equiv \sum_j^{\mathrm{occ.}} U_{ij}|\phi_j^{(\mathrm{eig})}\rangle. \tag{4.19}$$

Here we denote the Wannier states as $|\phi_i^{(\mathrm{WS})}\rangle$, which is different from the notations in Section 2.3. The density matrix is given by

$$\rho \equiv \sum_i^{\mathrm{occ.}} |\phi_i^{(\mathrm{WS})}\rangle\langle\phi_i^{(\mathrm{WS})}|. \tag{4.20}$$

The subsystems are given by

$$\rho_{\mathrm{A}} \equiv \sum_i^{\mathrm{A(occ.)}} |\phi_i^{(\mathrm{WS})}\rangle\langle\phi_i^{(\mathrm{WS})}|, \tag{4.21}$$

$$\rho_{\mathrm{B}} \equiv \sum_i^{\mathrm{B(occ.)}} |\phi_i^{(\mathrm{WS})}\rangle\langle\phi_i^{(\mathrm{WS})}| \equiv \rho - \rho_{\mathrm{A}}. \tag{4.22}$$

The number of states in the two subsystems are given in the forms of Eqs. (4.9) and (4.10), respectively.

Unlike eigen states, Wannier states do *not* form a subsystem ρ_{A} that commutes with the Hamiltonian :

$$0 \neq H\rho_{\mathrm{A}} - \rho_{\mathrm{A}}H. \tag{4.23}$$

This inequality is shown, when Eq. (4.19) is substituted into Eq. (4.21);

$$\begin{aligned}
\rho_{\mathrm{A}} &\equiv \sum_i^{\mathrm{A(occ.)}} |\phi_i^{(\mathrm{WS})}\rangle\langle\phi_i^{(\mathrm{WS})}| \\
&= \sum_i^{\mathrm{A(occ.)}} \sum_k^{\mathrm{occ.}} \sum_l^{\mathrm{occ.}} U_{ik}U_{il}^*|\phi_k^{(\mathrm{eig})}\rangle\langle\phi_l^{(\mathrm{eig})}| \\
&= \sum_k^{\mathrm{occ.}} \sum_l^{\mathrm{occ.}} \tilde{\delta}_{kl}|\phi_k^{(\mathrm{eig})}\rangle\langle\phi_l^{(\mathrm{eig})}|,
\end{aligned} \tag{4.24}$$

where

$$\tilde{\delta}_{kl} \equiv \sum_i^{\mathrm{A(occ.)}} U_{il}^*U_{ik} = \sum_i^{\mathrm{A(occ.)}} \left(U^{-1}\right)_{li} U_{ik}. \tag{4.25}$$

The matrix $\tilde{\delta}_{kl}$ is not the unit matrix ($\tilde{\delta}_{kl} \neq \delta_{kl}$), unless the 'A' subsystem contains *all* the occupied Wannier states.

From Eq. (4.23), Eq. (4.17) is not satisfied

$$0 \neq H_{\mathrm{map,I}}^{(\mathrm{A})}\rho_{\mathrm{A}} - \rho_{\mathrm{A}}H_{\mathrm{map,I}}^{(\mathrm{A})}. \tag{4.26}$$

In results, the formulation with $H_{\mathrm{map,I}}^{(\mathrm{A})}$ is not applicable to the subspace division with Wannier states, though applicable to the subspace division with eigen states.

General theory

Now we construct a general formulation that is applicable to both of the above two cases in the subspace division. We generally divide the system ρ into the subsystems $\rho_{\rm A}$ and $\rho_{\rm B}$ that satisfy

$$\rho = \rho_{\rm A} + \rho_{\rm B} \tag{4.27}$$
$$\rho_{\rm A}\,\rho_{\rm B} = 0. \tag{4.28}$$

The density matrix satisfies the commutation relation

$$0 = H\rho - \rho H. \tag{4.29}$$

It is useful to write down the following notation and relation

$$\bar{\rho} \equiv \hat{1} - \rho = \hat{1} - \rho_{\rm A} - \rho_{\rm B} \tag{4.30}$$
$$0 = H\bar{\rho} - \bar{\rho}H. \tag{4.31}$$

Here we propose another mapped Hamiltonian

$$H^{\rm (A)}_{\rm map,II} \equiv H + 2\eta_{\rm s}\rho_{\rm B} - \left(H\rho_{\rm B} + \rho_{\rm B}H\right), \tag{4.32}$$

which we call the mapped 'type II (two)' Hamiltonian. Using the energy shifted Hamiltonian

$$\Omega \equiv H - \eta_{\rm s}, \tag{4.33}$$

the mapped Hamiltonian is written as

$$H^{\rm (A)}_{\rm map,II} \equiv H - \Omega\rho_{\rm B} - \rho_{\rm B}\Omega. \tag{4.34}$$

We will investigate the condition of the following commutation relation

$$0 = H^{\rm (A)}_{\rm map,II}\rho_{\rm A} - \rho_{\rm A}H^{\rm (A)}_{\rm map,II}. \tag{4.35}$$

As a preparation, we obtain, from Eq. (4.28),

$$\begin{aligned}
\left[H\rho_{\rm B} + \rho_{\rm B}H,\ \rho_{\rm A}\right] &= \left(H\rho_{\rm B} + \rho_{\rm B}H\right)\rho_{\rm A} - \rho_{\rm A}\left(H\rho_{\rm B} + \rho_{\rm B}H\right) \\
&= \rho_{\rm B}H\rho_{\rm A} - \rho_{\rm A}H\rho_{\rm B}.
\end{aligned} \tag{4.36}$$

With the definition in Eq. (4.32), we calculate the commutator

$$\begin{aligned}
\left[H^{\rm (A)}_{\rm map,II},\ \rho_{\rm A}\right] &= [H,\rho_{\rm A}] + 2\eta_{\rm s}\left[\rho_{\rm B},\rho_{\rm A}\right] - \left[H\rho_{\rm B} + \rho_{\rm B}H,\ \rho_{\rm A}\right] \\
&= (H\rho_{\rm A} - \rho_{\rm A}H) + 0 - (\rho_{\rm B}H\rho_{\rm A} - \rho_{\rm A}H\rho_{\rm B}) \\
&= (1-\rho_{\rm B})H\rho_{\rm A} - \rho_{\rm A}H\,(1-\rho_{\rm B}) \\
&= (\bar{\rho}+\rho_{\rm A})H\rho_{\rm A} - \rho_{\rm A}H\,(\bar{\rho}+\rho_{\rm A}) \\
&= \bar{\rho}H\rho_{\rm A} - \rho_{\rm A}H\bar{\rho} \\
&= H\bar{\rho}\rho_{\rm A} - \rho_{\rm A}\bar{\rho}H,
\end{aligned} \tag{4.37}$$

where the second, fourth or last equality is obtained by Eq. (4.36), Eq. (4.30) or Eq. (4.31), respectively. The resultant quantity in Eq. (4.37), $(H\bar{\rho}\rho_{\rm A} - \rho_{\rm A}\bar{\rho}H)$, will be zero, if the commutation relations

$$[H, \rho_{\rm A}] = [H, \bar{\rho}] = 0 \tag{4.38}$$

$$[\bar{\rho}, \rho_{\rm A}] = 0 \tag{4.39}$$

are satisfied *or* the orthogonality relation

$$\bar{\rho}\rho_{\rm A} = 0 \tag{4.40}$$

is satisfied.

In the case of the subspace division with eigen states, in Eqs. (4.7),(4.8), the operator $\bar{\rho}$ is given as

$$\bar{\rho} \Rightarrow \sum_i (1 - f_i) |\phi_i^{(\rm eig)}\rangle\langle\phi_i^{(\rm eig)}| \tag{4.41}$$

and Eqs. (4.38),(4.39) are satisfied. Therefore, Eq. (4.35) is satisfied. In this case, we can directly prove Eq. (4.35), because Eq. (4.12) is satisfied and the commutation relation with the 'type II' mapped Hamiltonian, Eq. (4.35), is reduced to that with the 'type I' Hamiltonian, Eq. (4.17).

In the case of the subspace division with Wannier states, in Eqs. (4.21), (4.22), the operator $\bar{\rho}$ is given as

$$\bar{\rho} \Rightarrow \sum_i^{\rm unocc.} |\phi_i^{(\rm eig)}\rangle\langle\phi_i^{(\rm eig)}| \tag{4.42}$$

and Eq. (4.40) is satisfied. Therefore, Eq. (4.35) is satisfied.

We have proved that Eq. (4.35) is satisfied in the two cases of the subspace division with eigen states and Wannier states. The corresponding energy functional to be minimized is given as

$$\begin{aligned} E^{(\rm A)}_{\rm map,II}[\rho_A] &\equiv {\rm Tr}[H^{(\rm A)}_{\rm map,II}\rho_{\rm A}] \\ &= {\rm Tr}[H\rho_{\rm A}] + 2\eta_{\rm s}{\rm Tr}[\rho_{\rm A}\rho_{\rm B}] - ({\rm Tr}[\rho_{\rm B}H\rho_{\rm A}] + {\rm Tr}[\rho_{\rm A}H\rho_{\rm B}])\,. \end{aligned} \tag{4.43}$$

The second term is the 'penalty' term explained in Eq. (4.18). The expression of this term with eigen states is seen in Eq. (4.18). The expression with Wannier states is

$$2\eta_{\rm s}{\rm Tr}[\rho_{\rm A}\rho_{\rm B}] = 2\eta_{\rm s} \sum_i^{\rm A(occ.)} \sum_j^{\rm B(occ.)} |\langle\phi_i^{(\rm WS)}|\phi_j^{(\rm WS)}\rangle|^2. \tag{4.44}$$

As in Eq. (4.18), the minimization of this term results in the orthogonality relation with a sufficiently large value of $\eta_{\rm s}$. When the orthogonality is satisfied ($\rho_{\rm A}\rho_{\rm B} = 0$), the second term in Eq. (4.43) will vanish. The third term will also vanish, because of

$${\rm Tr}[\rho_{\rm B}H\rho_{\rm A}] = {\rm Tr}[\rho_{\rm A}\rho_{\rm B}H] = 0. \tag{4.45}$$

Note that the present theory is applicable to the subspace division with three or more subsystems $\rho_\alpha, \rho_\beta, \rho_\gamma \cdots$;

$$\rho = \rho_\alpha + \rho_\beta + \rho_\gamma \cdots \tag{4.46}$$

The density matrix of the 'α' subsystem ρ_α is determined by the above formulation with setting $\rho_A \Rightarrow \rho_\alpha$ and $\rho_B \Rightarrow \rho - \rho_\alpha = \rho_\beta + \rho_\gamma + \cdots$.

Theories in this chapter

In the following sections, we construct several theories for large-scale calculations using Eq. (4.35) as their foundation. In Section 4.2, the mean-field equation for a Wannier state is derived from Eq. (4.35). The formulation corresponds to a particular case of the subspace division in which the 'A' subsystem contains only one Wannier state. The derived mean-field equation will be solved approximately in the variational and perturbative order-N methods, described in Section 4.3. In Section 4.4, a hybrid scheme is given as a direct application of the energy functional of Eq. (4.43).

4.2 Mean-field equation for generalized Wannier states

Here a mean-field equation for the generalized Wannier states will be derived, using the formulation developed in the previous section.

Reformulation of generalized Wannier states

As in Section 2.3, the generalized Wannier states are defined as localized one-electron states that satisfy

$$H\phi_i = \sum_{j=1}^{N} \varepsilon_{ij}\phi_j, \tag{4.47}$$

where N is the number of occupied states. The parameters ε_{ij} are the Lagrange multipliers for the orthogonality constraints

$$\langle\phi_i|\phi_j\rangle = \delta_{ij} \tag{4.48}$$

and are given as

$$\varepsilon_{ji} = \langle\phi_i|H|\phi_j\rangle. \tag{4.49}$$

Before deriving the mean-field equation, we re-formulate Eqs. (4.47),(4.48), with the one-body density matrix

$$\rho \equiv \sum_{i=1}^{N} |\phi_i\rangle\langle\phi_i|. \tag{4.50}$$

The density matrix satisfies the commutation relation and the idempotency

$$\rho H = H\rho, \tag{4.51}$$

$$\rho^2 = \rho. \tag{4.52}$$

Now it is useful to define $H_{\rm occ}$ as

$$\rho H = H\rho = H_{\rm occ} \equiv \sum_{j=1}^{N} |\phi_j^{(\rm eig)}\rangle\varepsilon_j^{(\rm eig)}\langle\phi_j^{(\rm eig)}|, \tag{4.53}$$

which corresponds to the Hamiltonian projected on the occupied Hilbert space. With ρ and $H_{\rm occ}$, Eqs. (4.47) and (4.48) are rewritten as

$$(H - H_{\rm occ})\,|\phi_k\rangle = 0, \tag{4.54}$$

$$(\rho - 1)\,|\phi_k\rangle = 0, \tag{4.55}$$

respectively. The former equation is derived, when Eq. (4.49) is substituted into Eq. (4.47) :

$$\begin{aligned} 0 &= H|\phi_i\rangle - \sum_{j=1}^{N} \varepsilon_{ij}|\phi_j\rangle \\ &= H|\phi_i\rangle - \sum_{j=1}^{N} |\phi_j\rangle\langle\phi_j|H|\phi_i\rangle \\ &= (H - \rho H)|\phi_i\rangle \\ &= (H - H_{\rm occ})|\phi_i\rangle. \end{aligned} \tag{4.56}$$

Derivation of mean-field equation

In Section 4.1, an equation is given for a subsystem $\hat{\rho}_{\rm A}$ that is constructed from several selected Wannier states. Here we construct a subsystem which contains only one Wannier state $|\phi_i\rangle$;

$$\rho_{\rm A} \Rightarrow \rho_i \equiv |\phi_i\rangle\langle\phi_i| \tag{4.57}$$

$$\rho_{\rm B} \Rightarrow \bar{\rho}_i \equiv \sum_{j(\neq i)}^{\rm occ.} |\phi_j\rangle\langle\phi_j| = \rho - \rho_i \tag{4.58}$$

$$H^{\rm (A)}_{\rm map,II} \Rightarrow H^{\rm (i)}_{\rm WS} \equiv H - \Omega\bar{\rho}_i - \bar{\rho}_i\Omega \tag{4.59}$$

or

$$H^{\rm (i)}_{\rm WS} \equiv H + 2\eta_{\rm s}\bar{\rho}_i - H\bar{\rho}_i - \bar{\rho}_i H. \tag{4.60}$$

Here some notations are redefined. The commutation relation in Eq. (4.35) is rewritten as

$$0 = H^{\rm (i)}_{\rm WS}\rho_i - \rho_i H^{\rm (i)}_{\rm WS}. \tag{4.61}$$

Eq. (4.61) means that the selected Wannier state $|\phi_i\rangle$ is an eigen state of the mapped Hamiltonian $H^{\rm (i)}_{\rm WS}$;

$$H^{(i)}_{\rm WS}|\phi_i\rangle = \varepsilon^{(i)}_{\rm WS}|\phi_i\rangle, \tag{4.62}$$

This is the mean-field equation for the Wannier state $|\phi_i\rangle$ in the sense that the orthogonality constraint with the other Wannier states is included in the Hamiltonian $H^{(i)}_{\rm WS}$. The selected Wannier state is determined uniquely by Eq. (4.62) and the normalization constraint ($\langle\phi_i|\phi_i\rangle = 1$). For the solution of Eq. (4.62), the one-electron energy is given by

$$\varepsilon^{(i)}_{\rm WS} = \langle\phi_i|H^{(i)}_{\rm WS}|\phi_i\rangle = \langle\phi_i|H|\phi_i\rangle, \tag{4.63}$$

where the last equality is due to the orthogonality

$$\bar{\rho}_i|\phi_i\rangle = 0. \tag{4.64}$$

Now we prove that the mean-field equation (4.62) is equivalent to Eqs. (4.54) and (4.55);

$$\begin{aligned} 0 &= H^{(i)}_{\rm WS}|\phi_i\rangle - \varepsilon^{(i)}_{\rm WS}|\phi_i\rangle \\ &= H^{(i)}_{\rm WS}|\phi_i\rangle - |\phi_i\rangle\langle\phi_i|H|\phi_i\rangle \\ &= \left(H^{(i)}_{\rm WS} - \rho_i H\right)|\phi_i\rangle \\ &= 2\left(H - H_{\rm occ}\right)|\phi_i\rangle + 2\eta_{\rm s}(\rho - 1)|\phi_i\rangle, \end{aligned} \tag{4.65}$$

where the last equality is obtained by the operator equivalence

$$\begin{aligned} H^{(i)}_{\rm WS} - \rho_i H &= H + 2\eta_{\rm s}\bar{\rho}_i - H\bar{\rho}_i - \bar{\rho}_i H - \rho_i H \\ &= H + 2\eta_{\rm s}\left(\rho - \rho_i\right) - H\left(\rho - \rho_i\right) - \bar{\rho}_i H - \rho_i H \\ &= H + 2\eta_{\rm s}\left(\rho - \rho_i\right) - H\left(\rho - \rho_i\right) - \rho H \\ &= \left(H + H\rho_i - 2\rho H\right) + 2\eta_{\rm s}\left(\rho - \rho_i\right) \\ &= \left(H + H\rho_i - 2H_{\rm occ}\right) + 2\eta_{\rm s}\left(\rho - \rho_i\right) \end{aligned} \tag{4.66}$$

and the fact

$$\rho_i|\phi_i\rangle = |\phi_i\rangle\langle\phi_i|\phi_i\rangle = |\phi_i\rangle. \tag{4.67}$$

In short, a Wannier state $|\phi_i\rangle$ is not an eigen state of the original Hamiltonian H but an eigen state of the above mean-field Hamiltonian $H^{(i)}_{\rm WS}$. It should be noted that Eq. (4.62) is satisfied by Eqs. (4.54),(4.55) with an *arbitrary* choice of $\eta_{\rm s}$. So as to obtain the correct ground state with Eq. (4.62), the energy shift parameter $\eta_{\rm s}$ should be chosen as a sufficiently large value, which will be discussed below.

Another derivation of mean-field equation

The above mean-field equation (4.62) can be also derived in an alternative way [27]; The energy functional in Eq. (2.89) is rewritten, with the density matrix, as

$$\begin{aligned} E_{\rm O(N)} &= \sum_{i,j}^{N}(2\delta_{ij} - \langle\phi_j|\phi_i\rangle)\langle\phi_i|\Omega|\phi_j\rangle & (4.68)\\ &= {\rm Tr}[(2\rho - \rho^2)\Omega], & (4.69)\end{aligned}$$

where we denote the functional $E_{\rm O(N)}$, instead of the notation $E^{\rm (II)}_{\rm elec}$ used in Section 2.5. The variation with respect to one Wannier state $|\phi_i\rangle$ is given by

$$\begin{aligned} 0 &= \frac{\delta E_{\rm O(N)}}{\delta\langle\phi_i|}\\ &= (2\Omega - \rho\Omega - \Omega\rho)\,|\phi_i\rangle\\ &= (2H - \rho H - H\rho)\,|\phi_i\rangle + 2\eta_{\rm s}\,(\rho - 1)\,|\phi_i\rangle\\ &= 2\,(H - H_{\rm occ})\,|\phi_i\rangle + 2\eta_{\rm s}(\rho - 1)|\phi_i\rangle, & (4.70)\end{aligned}$$

which is equivalent to Eq. (4.65).

Analysis of mean-field Hamiltonian

Figure 4.1 shows an example of the density of state (DOS) of H and $H^{(i)}_{\rm WS}$ for silicon crystal [27]. Here the label i of a Wannier state $|\phi_i\rangle$ denote the bond site as its localization center. The DOS profile of the original Hamiltonian is decomposed into two parts, that is, the DOS profile of the valence band $D_{\rm val}(\varepsilon)$ and that of the conduction band $D_{\rm cond}(\varepsilon)$

$$D(\varepsilon) = D_{\rm val}(\varepsilon) + D_{\rm cond}(\varepsilon). \tag{4.71}$$

For the mean-field Hamiltonian $H^{(i)}_{\rm WS}$, the DOS profile $D^{(i)}_{\rm WS}(\varepsilon)$ is decomposed into three parts;

$$D^{(i)}_{\rm WS}(\varepsilon) = \delta(\varepsilon - \varepsilon^{(i)}_{\rm WS}) + D_{\rm cond}(\varepsilon) + D_{\rm high}(\varepsilon). \tag{4.72}$$

The first term is the isolated level shown in the figure. The second term is the same DOS profile as the conduction band in H. The last term is the high energy band located at $\varepsilon \geq 2\eta \approx 272$ eV. Hereafter we explain the three parts; The first term of Eq. (4.72) is the non-degenerate lowest eigen level of $H^{(i)}_{\rm WS}$ of which eigen state

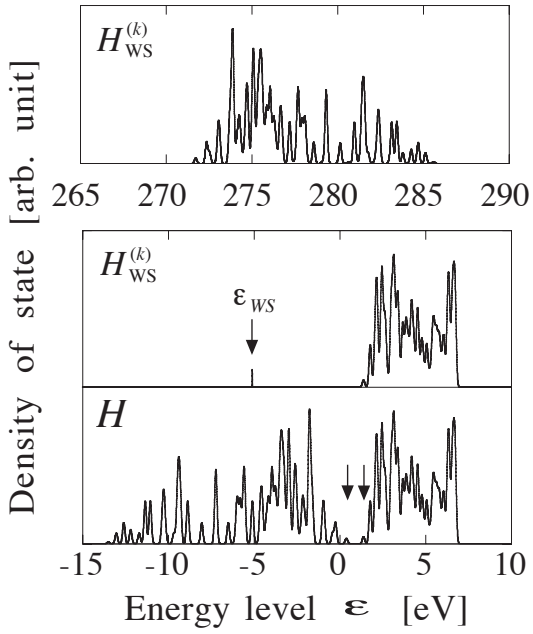


Figure 4.1: The DOS of Hamiltonians H and $H_{\rm WS}^{(k)}$ for the Si case with 512 atoms in the periodic cell [27]. The isolated level $\varepsilon_{\rm Wannierstate}$ is broadened by a Gaussian with a width 0.01eV and all the other levels are by that with a width of 0.1eV. The two parallel arrows in H indicate $\varepsilon_N^{(\rm eig)}$ and $\varepsilon_{N+1}^{(\rm eig)}$.

is the correct Wannier state $|\phi_i\rangle$. The value of $\varepsilon_{\rm WS}^{(i)}$ is the weighted center of the valence band, due to the following reason; Since all the bond sites are symmetrically equivalent in the diamond structure, the one-electron energy of the corresponding Wannier states should have a common value ($\varepsilon_{\rm WS}^{(i)} = \varepsilon_{\rm WS}$). Though the present calculation is done within the tight-binding formulation, the one-electron energy of Wannier state $\varepsilon_{\rm WS}$ is well defined, even within *ab initio* calculations, as the weighted band center of the valence band. On the other hand, the band structure energy is uniquely determined as the sum of the one-electron energies of the eigen states and the Wannier states. From the above requirement, the one-electron energy of the present Wannier state should be the weighted center of the valence band

$$\varepsilon_{\rm WS}^{(i)} = \varepsilon_{\rm WS} = \frac{1}{N}\sum_j^N \varepsilon_j^{(\rm eig)} = \frac{\int D_{\rm val}(\varepsilon)\varepsilon d\varepsilon}{\int D_{\rm val}(\varepsilon)d\varepsilon}. \tag{4.73}$$

The second term of Eq. (4.72) is contributed by the eigen states in the conduction band of H, ($\phi_i^{(\rm eig)}, N+1 \le i \le 2N$). They are also the eigen states of $H_{\rm WS}^{(k)}$ with the same eigen energies

$$H_{\rm WS}^{(k)}|\phi_i^{(\rm eig)}\rangle = H|\phi_i^{(\rm eig)}\rangle = \varepsilon_i^{(\rm eig)}|\phi_i^{(\rm eig)}\rangle. \tag{4.74}$$

The third term of Eq. (4.72) is contributed by all the other (N−1) occupied Wannier states ($\{\phi_j\}; j = 1, 2...., i-1, i+1, ..., N$). Such Wannier states are equivalent to $|\phi_i\rangle$ but their localization centers are different from the i-th bond site. They are *not*

eigen states of H nor $H_{\rm WS}^{(k)}$ but satisfy

$$\rho|\phi_j\rangle = |\phi_j\rangle \tag{4.75}$$

and

$$\langle\phi_j|H_{\rm WS}^{(i)}|\phi_l\rangle = 2\eta - \langle\phi_j|H_{\rm occ}|\phi_l\rangle, \quad j,l \neq i. \tag{4.76}$$

Due to Eq. (4.76), the corresponding band $D_{\rm high}(\varepsilon)$ has the property

$$D_{\rm high}(\varepsilon) \approx D_{\rm val}(2\eta - \varepsilon) \tag{4.77}$$

and is located at $\varepsilon \geq 2\eta - \varepsilon_N \approx 272$ eV. Due to the large energy shift of $D_{\rm high}(\varepsilon)$, the Hilbert space of the other Wannier states $(\{\phi_j\}; j = 1, 2...., i-1, i+1, ..., N)$ is automatically excluded in the variational freedom of ϕ_i, which results in the orthogonality constraint between the occupied Wannier states ($\langle\phi_i|\phi_j\rangle = \delta_{ij}$).

Locality as virtual impurity state

The locality of the Wannier state can be quantitatively discussed with the mean-field equation, as follows; Since the DOS profile of $H_{\rm WS}^{(i)}$ has one localized eigen level and a conduction band, the locality of the Wannier state ϕ_i is mapped formally to a virtual impurity state. The ionization energy is defined, also formally, as

$$\Delta_{\rm WS}^{(i)} \equiv \varepsilon_{N+1}^{\rm (eig)} - \varepsilon_{\rm WS}^{(i)} \tag{4.78}$$

Using the general uncertainty relation in the quantum mechanics, a length scale of the spatial spread is defined as

$$\xi_{\rm WS}^{(i)} \equiv \frac{\hbar}{\sqrt{2m_{\rm e}\Delta_{\rm WS}^{(i)}}}, \tag{4.79}$$

where $m_{\rm e}(\equiv 1{\rm a.u.})$ is the mass of electron. The above length should characterize the locality of ϕ_i as a virtual impurity state.

Here we explain that the length $\xi_{\rm WS}^{(i)}$ explains quantitatively the locality of the corresponding wave function $|\phi_i\rangle$. Before the practical silicon case, a simpler tight-binding Hamiltonian in the diamond structure is defined as follows:

$$H_0 = \sum_{i=1}^{N} \left(|{\rm b}_i\rangle\varepsilon_{\rm b}\langle{\rm b}_i| + |{\rm a}_i\rangle\varepsilon_{\rm a}\langle{\rm a}_i|\right). \tag{4.80}$$

Here $|{\rm b}_i\rangle$ or $|{\rm a}_i\rangle$ is the ideally sp^3 bonding or antibonding orbitals at the i-th bond site. The wave function $|{\rm b}_i\rangle$ and $|{\rm a}_i\rangle$ are completely localized on a pair of atoms on the$i$-th bond site. The energy difference between the bonding and antibonding orbitals is set to be $\varepsilon_{\rm a} - \varepsilon_{\rm b} = 8.25$eV, according to the present tight-binding Hamiltonian of silicon , Since the Wannier state is reduced to the bonding orbital ($|\phi_i\rangle \Rightarrow |{\rm b}_i\rangle$) and the lowest occupied level is that of the antibonding state ($\varepsilon_{N+1}^{\rm (eig)} \Rightarrow \varepsilon_{\rm a}$) , the ionization energy as the virtual impurity state is reduced to

$\Delta_{\rm WS}^{(i)} \Rightarrow \varepsilon_{\rm a} - \varepsilon_{\rm b} = 8.25$eV. Due to Eq. (4.79), the length scale for the spatial spread is calculated as

$$\xi_{\rm b} \equiv \frac{\hbar}{\sqrt{2m_{\rm e}(\varepsilon_{\rm a} - \varepsilon_{\rm b})}} = 0.29\, d_0, \tag{4.81}$$

where d_0 is the equilibrium bond length ($d_0 \equiv 2.4$ Å). The value of $\xi_{\rm b} = 0.29 d_0$ is consistent to the fact that the spread of a bonding orbital should be less than or about equal to the bond length d_0. Now we turn to the practical silicon case, in which $\Delta_{\rm WS} = 6.49$eV is obtained from Fig. 4.1. Using $\xi_{\rm b}$ as a reference, Eq. (4.79) give the value

$$\xi_{\rm WS}^{(i)} \equiv \frac{\hbar}{\sqrt{2m_{\rm e}\Delta_{\rm WS}^{(i)}}} = \xi_{\rm b}\sqrt{\frac{\varepsilon_{\rm a} - \varepsilon_{\rm b}}{\Delta_{\rm WS}^{(i)}}} = 1.13\, \xi_{\rm b}. \tag{4.82}$$

On the other hand, the spatial spread of the Wannier state can be directly estimated from the calculated wave function as

$$\bar{r}_{\rm WS}^{(i)} \equiv \sqrt{\langle \phi_i | (\hat{\boldsymbol{r}} - \boldsymbol{r}_i)^2 | \phi_i \rangle}, \tag{4.83}$$

where $\boldsymbol{r}_i \equiv \langle \phi_i | \hat{\boldsymbol{r}} | \phi_i \rangle$ is the localization center of the Wannier state. The actual values are calculated with the assumption that all atomic orbitals are localized at the atomic position. For a bonding orbital $|{\rm b}_i\rangle$, this length is the half of the bond length ($\bar{r}_{\rm b} \equiv d_0/2$) from its definition. For the practical Wannier state ϕ_i, we obtained

$$\bar{r}_{\rm WS}^{(i)} = 1.2\, \bar{r}_{\rm b}. \tag{4.84}$$

Here we can find that the ratio $\bar{r}_{\rm WS}^{(i)}/\bar{r}_{\rm b} = 1.2$ agrees with the corresponding value $\xi_{\rm WS}^{(i)}/\xi_{\rm b} = 1.13$ estimated from the uncertainty relation. Since the two ratios are calculated from different quantities, the agreement between them means that the locality of Wannier states is explained quantitatively as a virtual impurity state. A quantitative interpretation of the above results is that the Wannier state $|\phi_i\rangle$ has a slightly wider spatial spread than that of the bonding orbital $|{\rm b}_i\rangle$, due to the hoppings between bond sites.

It should be emphasized that the present discussion is based on an exact equation, Eq. (4.62). The locality of the Wannier state is derived from the eigen value distribution of the Hamiltonian $H_{\rm WS}^{(i)}$, which is independent on any explicit localization constraint.

4.3 Variational and perturbative order-N methods

The formulation in Section 4.2 gives a mean-field equation (4.62). The formulation is summarized here; The mean-field equation is given as

$$H_{\rm WS}^{(i)}|\phi_i\rangle = \varepsilon_{\rm WS}^{(i)}|\phi_i\rangle \tag{4.85}$$

where

$$\Omega \equiv H - \eta_{\rm s} \tag{4.86}$$

$$\bar{\rho}_i \equiv \sum_{j(\neq i)}^{\rm occ.} |\phi_j\rangle\langle\phi_j| \tag{4.87}$$

$$H_{\rm WS}^{\rm (i)} \equiv H - \Omega\bar{\rho}_i - \bar{\rho}_i\Omega \tag{4.88}$$

The corresponding energy functional to be minimized is given, by Eq. (4.69), as

$$E_{\rm O(N)} = {\rm Tr}[(2\rho - \rho^2)(H - \eta_{\rm s})] \tag{4.89}$$

with the density matrix

$$\rho = \sum_{i}^{\rm occ.} |\phi_j\rangle\langle\phi_i|. \tag{4.90}$$

The mean-field equation (4.85) directly gives practical order-N algorithms within variational or perturbative procedures.

Variational order-N method

From the mean-field equation, a variational procedure is constructed so as to generate approximate Wannier states. We call the above method 'variational order-N method' in this thesis. As the practical procedure, Eq.(4.85) is solved iteratively under the given localization constraint on each Wannier state

$$\{\phi_i\} \to \{H_{\rm WS}^{(i)}\} \to \{\phi_i\} \to \{H_{\rm WS}^{(i)}\} \to \cdots \tag{4.91}$$

If the localization constraint is relaxed, the calculation will be exact. Even with a given localization constraint, the variational order-N method has several choices for constructing the iterative algorithm, which will be discussed in Section 5.2.

Here we discuss the localization constraint in our program code. For each Wannier state ϕ_i, its localization center $\bar{\boldsymbol{r}}_i$

$$\bar{\boldsymbol{r}}_i \equiv \langle\phi_i|\hat{\boldsymbol{r}}|\phi_i\rangle \tag{4.92}$$

is calculated. The localization region for the i-th Wannier state is defined as the atoms within the cutoff radius from the localization center;

$$|\boldsymbol{R}_J - \bar{\boldsymbol{r}}_i| < r_i^{\rm (cut)}, \tag{4.93}$$

where $\boldsymbol{R}_J$ denotes the position of the J-th atom and $r_i^{(\mathrm{cut})}$ is the given cutoff radius. Since the localization center $\bar{\boldsymbol{r}}_i$ is automatically updated by Eq.(4.92) during the molecular dynamics, we should control only the cutoff radii $\{r_i^{(\mathrm{cut})}\}$. In our program code, the cutoff radius $r_i^{(\mathrm{cut})}$ can be dynamically controlled for each Wannier state during dynamical simulations, which will be discussed in Section 7.4.

A technical comment is added: Though the calculations in this thesis are done by the *spherical* cutoff, there might be more sophisticated method to define the localization region. For example, Wannier states at surface may have a wider spatial spread within the surface plane than that within the direction perpendicular to the surface, which is a suitable situation for a non-spherical localization region for Wannier states. Such a non-spherical localization constraint is one of future works.

Perturbative order-N method

Using the mean field equation (4.62), we can also construct a perturbative method to generate Wannier states. This method corresponds to a non-self-consistent solution of the mean field equation and we call the method 'perturbative order-N method'. The silicon crystal case is picked out, for example. As discussed in the previous section, the Wannier state of silicon crystal $|\phi_i\rangle$ is quite similar to the ideal sp^3 bonding orbital ($|\phi_i\rangle \approx |\mathrm{b}_i\rangle$) and the Hamiltonian is quite similar to the simpler one in Eq. (4.80) ($H \approx H_0$). Based on the above facts, a perturbative solution of the mean-field equation (4.62) is given by

$$|\phi_i^{(\mathrm{PT})}\rangle = C^{(0)}|\mathrm{b}_i\rangle + \sum_{j(\neq i)} C^{(j)}|\mathrm{a}_j\rangle \tag{4.94}$$

The coefficient $C^{(j)}$ is given by the standard first-order perturbation

$$\frac{C^{(j)}}{C^{(0)}} = \frac{\langle a_j|H|b_i\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}}. \tag{4.95}$$

The coefficient $C^{(0)}$ is determined by the normalization ($\langle\phi_i^{(\mathrm{PT})}|\phi_i^{(\mathrm{PT})}\rangle = 1$). In the perturbation terms, bonding orbitals without the i-th bond site ($\{|\mathrm{b}_j\rangle\}_{j\neq i}$) are 'excluded', because they are in the high-energy band in Fig.4.1. In a physical sense, a bonding orbital $|\mathrm{b}_j\rangle$ is occupied by the Wannier state $|\phi_j\rangle$, whose center is located on the bond site, and can not be occupied by the other Wannier state $|\phi_i\rangle$ ($i \neq j$). For the justification of the perturbative treatment, the contribution of the unperturbed term ($|C^{(0)}|^2$) should be almost one. In the present case of bulk silicon, the resultant value is $|C^{(0)}|^2 = 0.94$. Equations (4.94) and (4.95) are the foundation of the perturbative 'order-N' method. The details of the formulation and the analysis of the resultant wave functions are explained in Section 5.5 for silicon crystal and other diamond structure solids.

It is generally important that the formulation of Eqs. (4.94) and (4.95) does *not* include any explicit localization constraint on the wave functions. The locality of the wave function $|\phi_i^{(\mathrm{PT})}\rangle$ is given directly by the short-range property of the Hamiltonian H.

In the molecular dynamics simulations, we should calculate the total electronic energy $E_{\rm elec}$ and the corresponding force on each atom $\boldsymbol{F}_I$

$$E_{\rm elec} = \sum_i^{\rm occ.} \langle \phi_i^{\rm (PT)} | H | \phi_i^{\rm (PT)} \rangle \tag{4.96}$$

$$\boldsymbol{F}_I^{\rm (elec)} \equiv -\frac{\partial E_{\rm elec}}{\partial \boldsymbol{R}_I} = -\sum_i \langle \phi_i^{\rm (PT)} | \frac{\partial H}{\partial \boldsymbol{R}_I} | \phi_i^{\rm (PT)} \rangle, \tag{4.97}$$

where $\boldsymbol{R}_I$ is the position of the I-th atom.

Unlike the variational method, the perturbative order-N method does not contain any parameter to be controlled in dynamical simulations. Its technical details is limited to those in computational techniques, such as saving the required memory size, which will be discussed in Section 5.1. We will also discuss the parallelization of the perturbative order-N method in Section 5.4. Since the above procedures between different Wannier states are completely independent, the parallel computations can be done with respect to Wannier states.

Example in silicon crystal

As an example, silicon crystal is calculated within the perturbative order-N method using a standard work station. Examples with the variational order-N method will be shown later in this thesis. As explained in Chapter 1, Fig. 1.1 shows the CPU time per one MD step as the function of the number of atoms. The resultant CPU time shows a clear order-N property among $10^2 - 10^6$ atoms. In the program code, Eqs. (4.94), (4.95), (4.96), (4.97) are done explicitly, which dominates the total CPU time. We will discuss this point, in Section 5.4, with the parallelization of the program code. The electronic energy per one Wannier state $\varepsilon_{\rm WS} = \langle \phi_i | H | \phi_i \rangle$ was calculated and its deviation from the correct value was about 0.054 eV. The deviation corresponds to 1 % of the energy ($\varepsilon_{\rm WS}$) and to 10 % of the energy difference from that of the ideal sp^3 bonding orbital ($\varepsilon_{\rm b} - \varepsilon_{\rm WS}$). Here we can see that the present tight-binding Hamiltonian is a short-range operator and the value of its matrix element $\langle \phi_i | H | \phi_i \rangle$ is determined dominantly within a quite local area. Note that we will discuss, in Section 5.5, how the perturbative order-N method reproduces the energy systematically among the diamond structure solids.

The equilibrium lattice constant and elastic constants are also calculated using the perturbative Wannier states. The calculated lattice constant has an error of 2 % from the correct one [80]. The elastic constants are calculated within the deformed crystals [80]. Since the elastic constants are given by the second order perturbation of energy with respect to small deformations, they are expected to be reproduced by the first-order perturbation of the wave functions. Table 4.1 shows the results of the bulk modulus (B) and the shear moduli ($C_{11} - C_{12}$ and C_{44}). The results of the present work are compared with those in the exact diagonalization [6], an *ab initio* calculation [81], a classical model [82], and experiments. The error of the present order-N method from the exact diagonalization result is also shown inside the bracket, which is less than 10 %. The above discrepancy between the order-N method and the exact tight-binding calculation is not important, when these values are compared with *ab initio* or experimental values. Here a relatively large error is found in C_{44}, because the corresponding shear mode is inherently complicated due to

a rehybridization and an internal strain [81, 62]. A theoretical quantity $C_{44}^{(0)}$ is often defined as the shear modulus without relaxing the internal strain. See Appendix B.2 for details. The value of $C_{44}^{(0)}$ is larger than that of C_{44} from its definition ($C_{44}^{(0)} > C_{44}$). The calculated values of $C_{44}^{(0)}$ are 188.3, 198.5 or 111.0 GPa with the present work [80], the exact tight-binding calculation [6] or the *ab initio* calculation [81], respectively. Since a large discrepancy is found between the results of the exact tight-binding calculation and the *ab initio* calculation, the methodological problem originates mainly from the present tight-binding Hamiltonian, not from the perturbative order-N method. From above all, we can conclude that the perturbative order-N method gives satisfactory results of the elastic constants, at least, within the present tight-binding Hamiltonian.

Discussion

Here several discussions are made. The variational order-N method and the perturbative order-N method are based on the same mean-field equation. In the computational algorithm, the perturbative method is much simpler than the variational method, because the perturbative method does not contain a self-consistent loop. Now it may be interesting to derive classical models by further simplification of the perturbative method. Since the perturbative formulation gives the Wannier state $\{\phi_i^{(\mathrm{PT})}\}$ as explicit functions of the atomic coordinates $\{\boldsymbol{R}_I\}$, the resultant total energy in Eq. (4.96) gives a classical model, in the sense that the force on atoms

$$\frac{\partial E_{\mathrm{elec}}}{\partial \boldsymbol{R}_I} \tag{4.98}$$

can be calculated analytically. Though the first order perturbation is the simplest treatment within quantum mechanics, the resultant energy functional is still much more complicated than standard classical models with two- or three-body potentials. On the other hand, several physical quantities, such as elastic constants, can be well reproduced by simple classical models (See Appendix B.2). The above fact implies that further simplification of the present energy functional may be possible, if its applicability is limited to specific purposes. Such a derivation of classical models may be one of future works.

	Present work (error)	Exact TB	*ab initio*	classical	Exp.
B	82.3 (6.05 %)	87.6	93.0	101.4	97.8
C_{11} - C_{12}	92.3 (1.70 %)	93.9	98.0	75.0	101.2
C_{44}	97.9 (10.0 %)	89.0	85.0	56.4	79.6

Table 4.1: The elastic constants in the unit of GPa. The results are calculated by the following methods; (i) the present work with perturbative Wannier states [80], (ii) the exact diagonalization of the tight-binding Hamiltonian [6], (iii) the *ab initio* (LDA) calculation [81], (iv) the Stilinger-Weber potential [82], a standard classical model. The experimental values are also plotted. The present experimental values are from Ref. [6]. In the bracket of the present work shows the error from the value of the exact diagonalization.

4.4 Hybrid scheme by dividing Hilbert space

Based on the formulation in Section 4.1, another novel methodology, 'hybrid scheme', is given in this section. The essence of the present hybrid scheme is to divide the one-body density matrix into two parts, as in Section 4.1,

$$\rho = \rho_{\rm A} + \rho_{\rm B} \tag{4.99}$$

and calculate $\rho_{\rm A}$ and $\rho_{\rm B}$ by different methods. This method is important especially in inhomogeneous systems.

In this thesis, practical hybrid schemes are done using the subspace division with Wannier states. The density matrix is constructed from the occupied Wannier states ϕ_i as

$$\rho \equiv \sum_i^{\rm occ.} |\phi_i\rangle\langle\phi_i|. \tag{4.100}$$

Now the occupied Wannier states $\{\phi_i\}$ are decomposed into two groups 'A' and 'B'. The 'subsystems' $\rho_{\rm A}$ or $\rho_{\rm B}$ is constructed from the members of the group 'A' or 'B', respectively;

$$\rho_{\rm A} \equiv \sum_i^{\rm A(occ.)} |\phi_i\rangle\langle\phi_i|, \tag{4.101}$$

$$\rho_{\rm B} \equiv \sum_i^{\rm B(occ.)} |\phi_i\rangle\langle\phi_i|. \tag{4.102}$$

It should be noted that, though the subsystems $\rho_{\rm A}$ and $\rho_{\rm B}$ are formally given as the sums of the Wannier states, it is *not* necessary to calculate all the Wannier states. What we should calculate is the partial density matrices $\rho_{\rm A}$ and $\rho_{\rm B}$. In the example shown in this section, $\rho_{\rm B}$ will be explicitly constructed from the Wannier states, but $\rho_{\rm A}$ will not.

Practical procedures

We will demonstrate the formulation with a silicon crystal with surfaces. In this case, the inhomogeneous property stems from the difference of electronic states between the bulk and surface regions. In the bulk region, the Wannier states are characterized by sp^3 hybridized bonds and can be described by the perturbative order-N method, as explained in Section 4.3. These Wannier states $\{\phi_i\}$ in the bulk region are chosen as the members of the 'B' subsystem by Eq. (4.102). The other Wannier states belong to the 'A' subsystem. The subsystem $\rho_{\rm A}$ will should be given with the exact diagonalization method of the mapped Hamiltonian $H^{\rm (A)}_{\rm map}$ in Eq. (4.32). Since the perturbative order-N methods is a non-selfconsistent method, the algorithm of the hybrid scheme is a 'one-way' algorithm; We first generate ρ_B using the perturbative order-N method, without ρ_A, and then we determine ρ_A with the given ρ_B. Such 'one-way' algorithm is of great advantage in saving the computational costs.

The present demonstration is done with a silicon crystal with an *unreconstructed* (001) surface. Due to the *unreconstructed* surface, the system is unstable and the

electronic energy gap almost vanishes (0.025 eV), as explained below. Therefore, the present example is one of the severest tests for the present methodology.

The actual sample is a slab with 16 atomic layers. Each atomic layers contains 64 atoms. The total number of atoms is $8 \times 8 \times 16 = 1024$ in the periodic cell. The z coordinate is written in the length unit of the interval between atomic layers; that is, the atoms lie in $z = 0, 1, 2....15$. The surface atoms are localized as $z = 0$ and have dangling bonds. In a physical sense, the value of z corresponds to the depth from the surface. The atoms on the opposite boundary surface of the sample ($z = 15$) do not have any dangling bond and is terminated by the Wannier states in the ideally sp^3 bonding orbitals. The opposite surface does not form a surface band and will be ignored in the below physical discussions. The electronic system in the calculation code contains $N \equiv 1984$ occupied states without the terminated Wannier states. The corresponding number of electrons is $2N$ with the para-spin factor (two). The center of Wannier state is denoted as

$$z_i \equiv \langle \phi_i^{(\mathrm{PT})} | z | \phi_i^{(\mathrm{PT})} \rangle. \tag{4.103}$$

Except the surface states, Wannier states shows bonding characters in the diamond structure. Each bond is placed among successive two layers and the center z_i of the corresponding Wannier state is located at half integers in the present unit ($z_i = 0.5, 1.5, 2.5...., 14.5$). The subsystem 'B' is constructed from the perturbative Wannier states whose centers are located at atomic layers deeper than the $z^{(c)}$-th layer ($z_i > z^{(c)}$)

$$\rho_{\mathrm{B}} \equiv \sum_{i}^{z_i > z^{(c)}} |\phi_i^{(\mathrm{PT})}\rangle\langle\phi_i^{(\mathrm{PT})}|. \tag{4.104}$$

The value of $z^{(c)} = 8$ is chosen. The other electronic states that correspond to Wannier states near the surface regions ($z_i < z^{(c)}$), belong to the 'A' subsystem. The divided subsystems 'A' and 'B' contain $N_A = 1088$ and $N_B = N - N_A = 896$ electronic states, respectively. The energy shift parameter is chosen as $2\eta_{\mathrm{s}} = 1$a.u. (about 27.2 eV). With the above preparations, the mapped Hamiltonian for the 'A' subsystem (ρ_{A}) is well defined as

$$H_{\mathrm{map}}^{(\mathrm{A})} \equiv H + 2\eta_{\mathrm{s}}\rho_{\mathrm{B}} - (H\rho_{\mathrm{B}} + \rho_{\mathrm{B}}H)\,. \tag{4.105}$$

The Hamiltonian $H_{\mathrm{map}}^{(\mathrm{A})}$ corresponds to $H_{\mathrm{map,II}}^{(\mathrm{A})}$ in Section 4.1, that is, here we drop the suffix 'II'.

As an additional approximation, we ignore the third term $(H\rho_{\mathrm{B}} + \rho_{\mathrm{B}}H)$ in the above formulation, due to the following reason; In the bulk region, the present (perturbative) Wannier state $|\phi_i\rangle$ is quite similar to the ideal sp^3 bonding state $|b_i\rangle$ at one bond sites ($|\phi_i\rangle \approx |b_i\rangle$), The Hamiltonian H, on the other hand, is quite similar to that in Eq. (4.80) ($H \approx H_0$), whose eigen states are $|b_i\rangle$ ($H_0|b_i\rangle = \varepsilon_{\mathrm{b}}|b_i\rangle$). Within the above approximation,we obtain

$$H\rho_{\mathrm{B}} + \rho_{\mathrm{B}}H \approx 2\varepsilon_{\mathrm{b}}\rho_{\mathrm{B}}. \tag{4.106}$$

and Eq.(4.105) is reduced to

$$H_{\mathrm{map}}^{(\mathrm{A})} \approx H + 2(\eta_{\mathrm{s}} - \varepsilon_{\mathrm{b}})\rho_{\mathrm{B}}. \tag{4.107}$$

When $\eta_{\rm s}$ is redefined, Eq. (4.107) is equivalent to Eq. (4.105) without the third term. It is noteworthy that in the context of Section 4.1, the above simplification means the reduction of the 'type II' Hamiltonian, given in Eq. (4.32), into the 'type I' Hamiltonian, given in Eq. (4.16). Off course, the above simplification is not necessary, in principle. In other words, we will demonstrate a simplest formulation of the hybrid scheme.

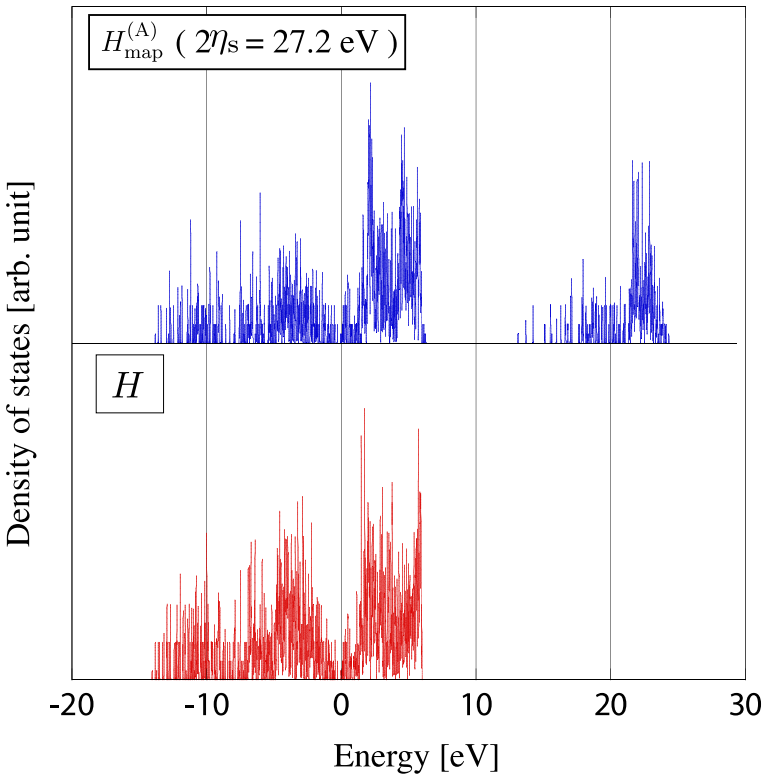


Figure 4.2: The density of states (DOS) in the hybrid method. Lower panel : DOS of the original Hamiltonian H. Upper panel : DOS of the mapped Hamiltonian $H^{(\rm A)}_{\rm map}$. The system is a silicon crystal with an unreconstructed surface.

Result

Figure 4.2 shows the result of the hybrid scheme. Among the DOS profiles in the present section, the energy origin ($\varepsilon = 0$) is shifted downward by 0.642 eV for the eye guide. The lower panel of Fig. 4.2 shows the DOS profile of the original Hamiltonian H. The highest occupied level ($\varepsilon_{\rm HO}$) and the lowest unoccupied level ($\varepsilon_{\rm LU}$) are given as $\varepsilon_{\rm HO} = -0.0236$ eV and $\varepsilon_{\rm LU} = 0.00048$ eV, respectively. As already explained, the calculated electronic energy gap, $\varepsilon_{\rm LU} - \varepsilon_{\rm HO} \approx 0.025$ eV, is much smaller than that of the bulk system (1.2 eV), due to the presence of the unreconstructed surface. The upper panel of Fig. 4.2 shows the DOS profile of the 'A' subsystem ($H^{(\rm A)}_{\rm map}$), which is well decomposed by three bands: The lower band ($\varepsilon < 0$) is that of $\rho_{\rm A}$ with N_A states. The exact value of its band top is $\varepsilon^{(A)}_{\rm top} \equiv -0.0539$ eV. The middle band ($0{\rm eV} < \varepsilon < 7{\rm eV}$) is the unoccupied band. The exact value of its band bottom

is $\varepsilon_{\rm bot}^{(U)} \equiv -0.00015$ eV . The upper band (13eV $< \varepsilon <$ 25eV) is that of $\rho_{\rm B}$ with N_B states. From above observations, the Hilbert space of the 'B' subsystem (ρ_B) is separated in the eigen value distributions of H_A, since the corresponding band is shifted in the energy domain near the value of $2\eta_{\rm s} = 27.2$eV. As details, two points are also discussed; First, the bottom of the unoccupied band in H ($\varepsilon_{\rm LU}$) is equal to the bottom of the middle band in H_A ($\varepsilon_{\rm bot}^{(U)}$) with a negligible error (less than 1 meV). This property will be exactly satisfied, if $\rho_{\rm B}$ is constructed from the exact Wannier states. Second, the top of the lower band in H_A ($\varepsilon_{\rm top}^{(A)} = -0.054$ eV) is close to the top of the occupied band in H ($\varepsilon_{\rm HO} = -0.024$ eV). This property is expected, though not exactly satisfied, if $\rho_{\rm B}$ is constructed from the exact Wannier states, because the top of the occupied band in H should be contributed dominantly by the Wannier states near the surface regions. The above two points imply that the present hybrid scheme can describe the surface band, which is essential to the reconstruction processes.

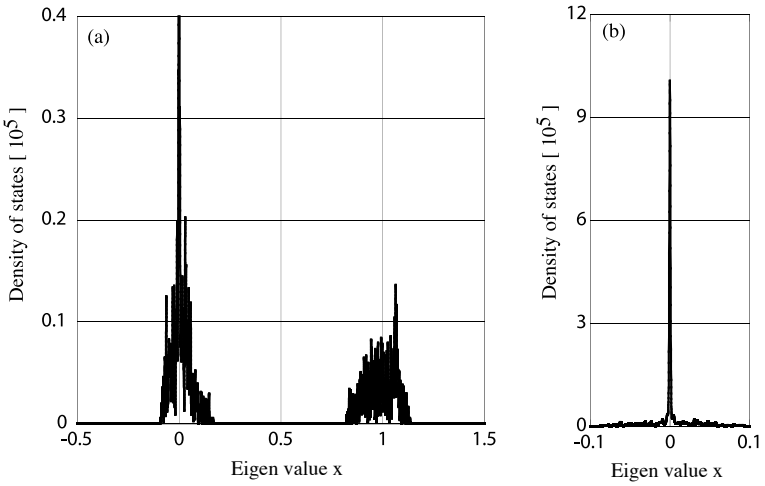


Figure 4.3: The density of states (DOS) of the partial density matrix ($\rho_{\rm B}$) is plotted with different ranges in (a) and (b). Note that the present DOS profile is given by the Gaussian broadening of each eigenvalues x with the broadening width of $\Delta x = 0.01$

Fig. 4.3 shows the DOS of the density matrix $\rho_{\rm B}$. If the density matrix $\rho_{\rm B}$ is constructed by exact orthogonal wave functions, the density matrix will show the exact idempotency ($\rho_B^2 = \rho_B$) and their eigen values x will be one for the N_B states that contribute $\rho_{\rm B}$ and zero for the other states. The calculated DOS using the perturbative Wannier states deviates from the exact idempotency but we can observe that the eigen values are well separated by the region of $x \approx 0$ and $x \approx 1$. The band of $x \approx 1$ contains N_B states. From the above facts, the present density matrix ρ_B shows a satisfactory property in the approximate idempotency. Here we discuss the details of the observed DOS profile in Fig. 4.3(a). The DOS profile can be decomposed by three parts; Two of them are the bands near the region of $x \approx 0$ and $x \approx 1$ that have finite band widths. The other one is the ideal delta function at $x = 0$, which is clearly seen on Fig. 4.3 (b). The bands with finite band widths originated from the fact that the perturbative Wannier states deviates from the exact

orthogonality ($\langle\phi_i^{(\mathrm{PT})}|\phi_j^{(\mathrm{PT})}\rangle \neq \delta_i$). The ideal delta function at $x = 0$ originates from the fact that the perturbative Wannier states in the 'B' subsystem are located only within the region of $z_i > z^{(c)}$. The resultant density matrix ρ_{B} has its spatial spread within the region of $z \geq z^{(B)} \equiv z^{(c)} - 2$, because each perturbative Wannier state has its spatial spread within the *second* nearest neighbor bond sites. Within the region of $z < z^{(B)}$, layers near the surface, the corresponding matrix elements in ρ_{B} are exactly zero both in the diagonal and off-diagonal elements, which forms the ideal delta functions on the DOS.

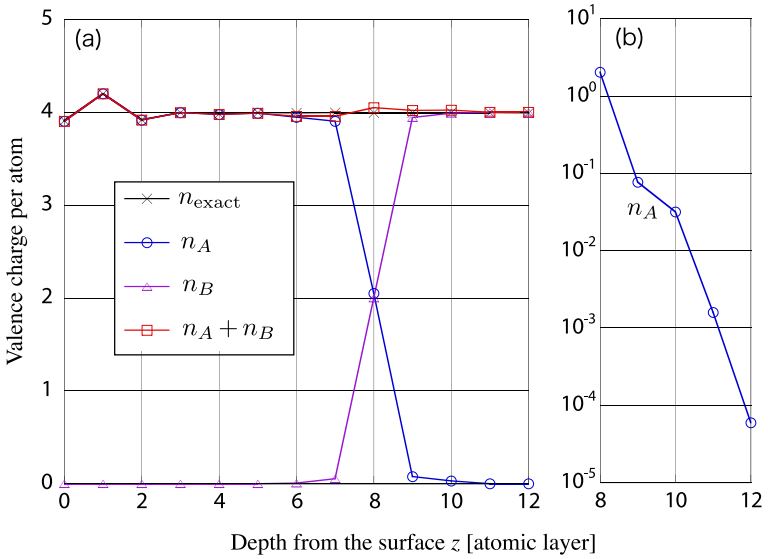


Figure 4.4: The averaged valence charge per atom with the function of the depth from the surface (z) in the unit of the atomic layer. The atoms at $z = 0$ correspond to the surface atoms.

Discussion

The real space picture of the present hybrid scheme is discussed. In Fig. 4.4(a), the averaged valence charge per atom $n(z)$ is plotted as the function of the atomic coordinate z in the length unit of the interval between atomic layers. The result with the standard diagonalization method is plotted as $n_{\mathrm{exact}}(z)$, which shows small deviations from the neutrality ($n = 4$) near the surface regions. Note that the surface atoms in the present sample are in the ideal crystalline geometry and will be reconstructed to form asymmetric dimers of partially ionic atoms, as discussed later in this thesis. The charge of the subsystems are also plotted as $n_A(z)$ and $n_B(z)$ and are defined by the diagonal elements of the density matrices ρ_A and ρ_B, respectively. In Fig. 4.4(a), we can observe that the total charge in the hybrid scheme ($n_A + n_B$) reproduces well that of the exact calculations (n_{exact}). The maximum error is only 1 % at $z = z^{(c)} = 8$ and the error at the surface atom ($z = 0$) is less than 0.2 %. The charge distributions of the two subsystems are overlapped in real space. For example, the charge at $z = z^{(c)} \equiv 8$, is contributed by n_A and n_B with almost same weights. This can be understood within the Wannier state picture as follows; Each Wannier state, except the surface one, is quite similar to an sp^3 bonding state on a bond site.

A Wannier state whose bond site is placed between the 8-th and 9-th atomic layers belongs to the 'B' subsystem and has almost the half of the weight on the 8-th atomic layer. On the other hand, a Wannier state whose bond site is placed between the 7-th and 8-th atomic layers belongs to the 'A' subsystem and has almost the half of the weight on the 8-th atomic layer. Figure 4.4(b) shows that the charge distribution of the 'A' subsystem $n_A(z)$ decays quickly at $z > 8$, which corresponds to the 'tail' of the Wannier states that belong to the 'A' subsystem. Here we note the following two issues; (i) The present hybrid scheme does not require any explicit constraint on the charge distribution $n(z)$ for its local charge neutrality ($n(z) \approx 4$) in the bulk regions. (ii) Though the above explanations are done within the Wannier state picture, the 'A' subsystem (ρ_A) is constructed from eigen states in the computational code. The present hybrid scheme does not require any explicit constraint on eigen states so as to decay in the region of $z > 8$. The above two properties are fulfilled only by the orthogonality between the two subsystems. In results, the two subsystems are overlapped in real space but orthogonal in the Hilbert space. This is the reason why we call the present method as 'division in Hilbert space'.

Hybrid scheme between order-N methods

Since the present hybrid scheme gives a well-defined mapped Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$, we can use, in principle, any quantum mechanical method for calculating the partial density matrix ρ_{A}. So far the hybrid scheme is done by the combination of the exact diagonalization method and the perturbative order-N method. In Chapter 7, several fracture simulations are done in the hybrid scheme of the variational order-N method and the perturbative order-N method. Another example is seen in Fig. 4.5 [52], in which we use the hybrid scheme of the recursion method (See Section 2.5) and the perturbative order-N method. The result in Fig. 4.5 should be compared with that by the exact diagonalization method, in the upper panel of Fig. 4.2.

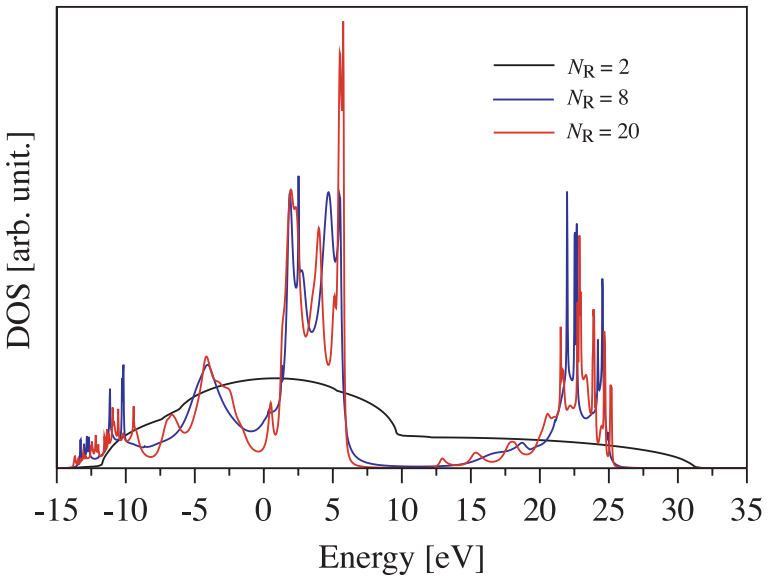


Figure 4.5: The density of states (DOS) in the hybrid method [52]. DOS for $H^{\rm A}_{\rm map}$ is calculated by the recursion method with the recursion order of $N_{\rm R} = 2, 8, 20$. The system is a silicon crystal with an unreconstructed surface. The exact diagonalization result is given in the upper panel of Fig. 4.2 for the same mapped Hamiltonian $H^{\rm (A)}_{\rm map}$.

4.5 Summary and future aspects

In this chapter, we constructed the theories of large-scale electronic structure calculations. The overview is shown in Fig. 4.6.

We derived a general equation of the partial density matrix, Eq. (4.35), in the form of a commutation relation (Section 4.1). As a specific case of the above equation, we derived the mean-field equation, Eq. (4.62) , of the generalized Wannier states (Section 4.2). These are the exact equations. As practical order-N methods, the variational and perturbative methods were constructed as approximate solutions of the above mean-field equation (Section 4.3). From Eq. (4.35), we also constructed the hybrid scheme by dividing the occupied Hilbert space. In the hybrid scheme, the electron system is divided into several subsystems in the Hilbert space and the subsystems are solved by different methods (Section 4.4). We prepared the program code of the hybrid scheme between the following methods; (i) the exact diagonalization method and the perturbative order-N method, (ii) the variational order-N method and the perturbative order-N method, (iii) the recursion method and the perturbative order-N method. The cases (i) and (iii) were demonstrated in Section 4.4. The case (ii) will be used in the practical large-scale fracture simulations in Chapter 7.

For the above theories, several details and applications will be discussed in the next chapter (Chapter 5). Particularly, the perturbative order-N method will be parallelized in Section 5.4. In Section 7.4, we will discuss several technical details related to the dynamical simulations.

Hereafter we discuss the future aspects of the theories. Since all the theories are well defined in quantum mechanics, the future aspect should be discussed in practical points;

(I) The first point is the parallelization of the variational order-N method. This point will be discussed in Section 7.7.

(II) The second point is the further application of the perturbative order-N method. In this thesis, the application is limited to the sp^3-hybridized bond. Since the mean-field equation of the Wannier state is a general equation, the perturbative Wannier states can be always constructed, if reliable unperturbed wave functions are prepared.

(III) The third point is the further application of the hybrid scheme. In the present thesis, the practical large-scale application is done by the hybrid scheme between the variational and perturbative order-N methods. Since the hybrid scheme is based on a general division principle in the Hilbert space, applications with other methods are straightforward. Now we are developing the program code of the recursion method, which was tested in a hybrid scheme (See Fig. 4.5).

It should be noted that, since the perturbative or variational order-N method is based on the Wannier state, their success is limited to covalent bonding materials. The recursion method, on the other hand, is not based on the generalized Wannier states and will be applied, especially, to metals.

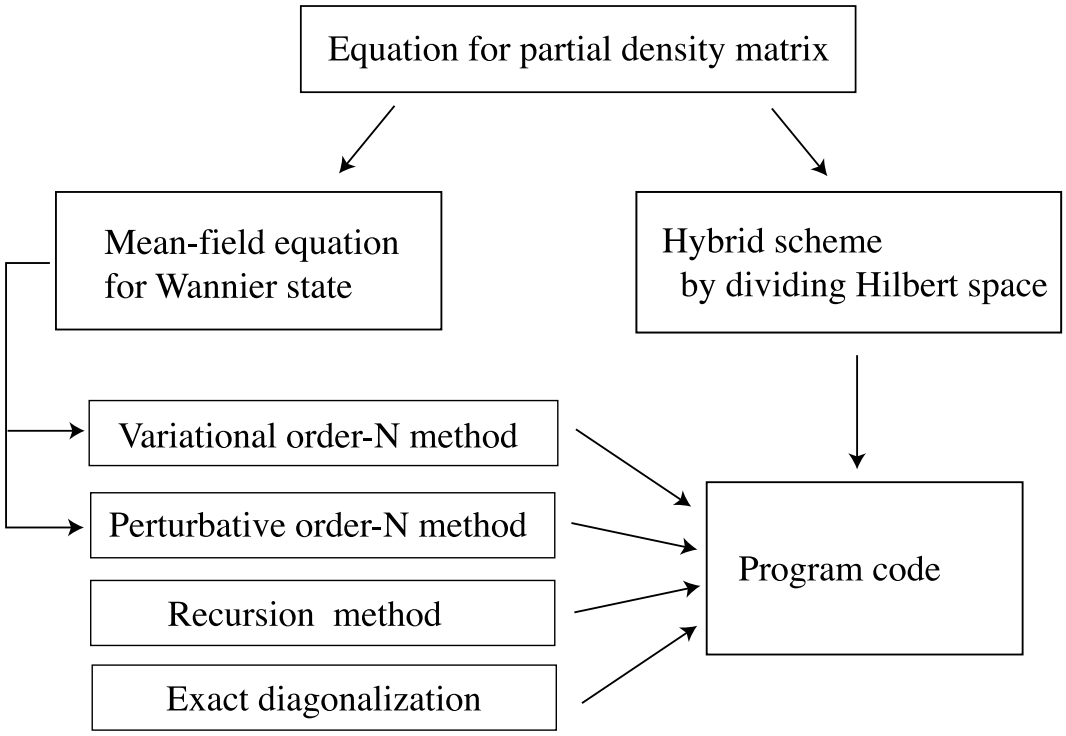


Figure 4.6: Overview of the present theories for large-scale electronic structure calculations.

Chapter 5

Details and applications

5.1 Perturbative order-N method

In this section, we discuss several technical details of our program code. Within the perturbative order-N method, the required operation number to generate the Wannier state is uniquely determined. In the practical computations, however, the optimal program code may differ among the hardware environments.

Limitation in memory size

A crucial problem is the size limitation of the built-in memory. For example, the present calculation is done by a standard work station with 2 GB built-in memory. In the perturbative order-N method, the calculation of the non-zero Hamiltonian matrix elements ($\{H_{\alpha I \beta J}\}$) is inevitable. For a system with $N_{\rm A}$ atoms, the number of the non-zero matrix elements is estimated to be $\nu^2 N_{\rm int} N_{\rm A}$, where ν is the number of orbitals per atom and $N_{\rm int}$ is the number of neighbor atoms within the interaction range. In the present Hamiltonian, $\nu = 4$ and $N_{\rm int} \approx 20$. If a matrix element is stored in the memory as a real value with the double precision, it costs eight byte (8B) in the memory. In the system with one million atoms ($N_{\rm A} = 10^6$), the total requirement of the memory size is

$$8[\mathrm{B}] \times \nu^2 \times N_{\rm int} \times N_{\rm A} \approx 8[\mathrm{B}] \times 16 \times 20 \times 10^6 \approx 2.5[\mathrm{GB}], \tag{5.1}$$

which will exceed the limitation of the present work station (2 GB). One simple solution for the memory size saving is not to store the Hamiltonian matrix in the memory. Since the present tight-binding Hamiltonian is an explicit function of the atomic coordinates ($H_{I\alpha,J\beta} \equiv H_{\alpha,\beta}(\boldsymbol{R}_I - \boldsymbol{R}_J)$), their values can be always calculated when the atomic coordinates $\{\boldsymbol{R}_I\}$ are given. The data size of the atomic coordinates $\{\boldsymbol{R}_I\} \equiv \{X_I, Y_I, Z_I\}$ is given as

$$8[\mathrm{B}] \times 3 \times N_{\rm A} \approx 8[\mathrm{B}] \times 3 \times 10^6 \approx 24[\mathrm{MB}], \tag{5.2}$$

which is quite smaller than that in Eq. (5.1). So as to generate one Wannier state $|\phi_i^{(\mathrm{PT})}\rangle$, only a partial Hamiltonian matrix is required as a work array. The size of the work array is determined by the spatial spread of the perturbative Wannier states $|\phi_i^{(\mathrm{PT})}\rangle$ in Eq. (4.94). Since the perturbative Wannier state contains about $N_{\rm int}$ atoms in its spatial spread, the work array for the partial Hamiltonian matrix requires the memory size of

$$8[\mathrm{B}] \times \nu^2 \times N_{\rm int} \times N_{\rm int} \approx 8[\mathrm{B}] \times 16 \times 20 \times 20 \approx 50[\mathrm{KB}], \tag{5.3}$$

which is negligible in the present hardware environment. The resultant Wannier state $|\phi_i^{(\mathrm{PT})}\rangle$ should be also stored in another work array, because the total memory requirement for storing all the perturbative Wannier states is

$$8[\mathrm{B}] \times \nu \times N_{\rm int} \times 2 \times N_{\rm A} \approx 8[\mathrm{B}] \times 4 \times 20 \times 2 \times 10^6 \approx 1.3[\mathrm{GB}] \tag{5.4}$$

with $2N_{\rm A}$ (doubly occupied) Wannier states in silicon. A work array for a Wannier state $|\phi_i^{(\mathrm{PT})}\rangle$, on the other hand, requires only a negligible memory size of

$$8[\mathrm{B}] \times \nu \times N_{\rm int} \approx 8[\mathrm{B}] \times 4 \times 20 \approx 0.6[\mathrm{KB}]. \tag{5.5}$$

The use of work arrays does not cause any numerical error, because the procedures among the different Wannier states are independent, as Eqs. (4.96),(4.97).

Here we note that the reduction of the memory requirement from Eqs. (5.1),(5.4) into Eqs. (5.3),(5.5) means the reduction from the memory size in an $O(N)$ system-size scaling into that in an $O(N^0) = O(1)$ scaling. The memory size in an $O(N)$ scaling is required with *classical* freedoms, say the atomic coordinates, as in the classical molecular dynamics.

'Domain' method as an optimal algorithm

In the above algorithm with work arrays, however, each Hamiltonian matrix element should be multiply calculated among the procedures of different Wannier states, because the localization regions for Wannier states are fairly overlapped in real space. Such multiple calculations of the same quantities cause an extra overhead in the CPU time. In short, the above algorithm saves the memory size but wastes the CPU time.

Fortunately, we can construct an algorithm that is advantageous for saving *both* in the CPU time and the memory size. In essence, we divided the system into some 'domains' in real space. We call the method as 'domain method'. We explain the algorithm, for simplicity, in the case of a one-dimensional (chain) system, in which the i-th Wannier state has its localization center on the i-th bond site. The i-th bond site lies between the $i-1$-th and the $i+1$-th bond sites. We classify the Wannier states into some real space 'domains'; the first domain contains the Wannier states of $\{\phi_1, \phi_2, \phi_3, ...\phi_{10}\}$ and the second domain contains those of $\{\phi_{11}, \phi_{12}, \phi_{13}, ...\phi_{20}\}$, and so on. The Hamiltonian matrix is stored as a work array for each domain. The work array should store all the Hamiltonian matrix elements that are required to generate all the Wannier states in the domain. The data in the work array are commonly used among all the Wannier states in the domain. In this domain method, a Hamiltonian matrix element will be multiply calculated, when it is included in several domains. For example, some of the atoms that are contained in ϕ_{10} may be included in the first and second domains. The corresponding matrix elements should be calculated both in the procedures of the first and second domains, which will results only in a small additional cost in the CPU time. In the view point of the memory size, the number of elements in the work array is $\nu^2 N_{\rm int} N_{\rm A}^{\rm (dom)}$, with the number of atoms per domain ($N_{\rm A}^{\rm (dom)}$). In the present case with silicon crystal, the domains are defined in the three dimensional space and each domain contains about $N_{\rm A}^{\rm (dom)} \approx 4000$ atoms. The requirement in the memory size is estimated to be

$$8[\mathrm{B}] \times \nu^2 N_{\rm int} N_{\rm A}^{\rm (dom)} \approx 8[\mathrm{B}] \times 16 \times 20 \times 4000 \approx 10[\mathrm{MB}], \tag{5.6}$$

which consumes only a small part of the built-in memory (2 GB) in the present work station. In the above example, the 'domain' method saves about 50 % of the actual computational time. For the optimal performance, the value of $N_{\rm A}^{\rm (dom)}$ should be chosen to be as large as possible within the built-in memory size. The use of the 'domain' method dose not cause any numerical error.

Several comments are added on the domain method; (i) Though the domain method requires an additional memory cost of the size with Eq. (5.6), the size is

still in a $O(N^0) = O(1)$ system-size scaling. (ii) The efficiency of the 'domain' method may depend on details of the hardware environments, such as the access speed of the memory, the size of the cache memory, the compiler options, and so on. (iii) In the domain method, the procedures between different domains are completely independent, which is the foundation of parallel computations.

5.2 Variational order-N method

The variational order-N method was explained in Section 4.3. This section gives its technical details. Particularly, we explain the algorithm for generating the wave functions with given atomic coordinates, which corresponds to the algorithm within one time step in molecular dynamics simulation. The technical details in dynamical simulations will be discussed in Section 7.4.

Iterative algorithm for self-consistent loop

Since the Wannier states should satisfy the mean-field equation of Eq. (4.85), the variational order-N method needs an iterative procedure for the self-consistency. The iterative loop was schematically shown in Eq. (4.91) and its details are described here. The deviation from the correct solution, Eq. (4.70), is given by

$$\begin{aligned} \delta\phi_i &\equiv \frac{\delta E_{\mathrm{O(N)}}}{\delta\langle\phi_i|} \\ &= (2\Omega - \rho\Omega - \Omega\rho)|\phi_i\rangle, \end{aligned} \tag{5.7}$$

which will decrease iteratively. We monitor the norm of the vector ($|\delta\phi_i|$) as a measure of the convergence to the correct solution. Due to the localization constraint, the deviation $|\delta\phi_i|$ can not reach to the exact zero. In the program code, we will quit the iterative loop, when the deviation $|\delta\phi_i|$ stops the exponential decrease.

Here we explain one of the iterative loop for updating the wave functions:

$$\{|\phi_i^{(\mathrm{old})}\rangle\} \Rightarrow \{|\phi_i^{(\mathrm{new})}\rangle\}. \tag{5.8}$$

(i) The one-body density matrix is constructed

$$\rho = \sum_i^N |\phi_i\rangle\langle\phi_i|. \tag{5.9}$$

(ii) The following matrix is constructed as a preparation

$$\tilde{H}_{\mathrm{WS}} \equiv H - \rho\Omega - \Omega\rho. \tag{5.10}$$

(iii) For each Wannier state ϕ_i, the mean-field Hamiltonian $H_{\mathrm{WS}}^{(i)}$ is constructed within its localization region

$$H_{\mathrm{WS}}^{(i)} = \tilde{H}_{\mathrm{WS}} + |\phi_i\rangle\langle\phi_i|\Omega + \Omega|\phi_i\rangle\langle\phi_i|. \tag{5.11}$$

(iv) For each mean-field Hamiltonian $H_{\mathrm{WS}}^{(i)}$, Eq. (4.85) is solved numerically. The lowest eigen state is obtained by the Lanczos method, which is reviewed in Appendix D.3. The initial vector for the Lanczos series is chosen as the old wave function $|\phi_i^{(\mathrm{old})}\rangle$. The number of the Lanczos series n_{L} is chosen, typically, $n_{\mathrm{L}} = 10$;

$$H_{\mathrm{WS}}^{(i)},\ |\phi_i^{(\mathrm{old})}\rangle \Rightarrow (\text{Lanczos method for Eq. (4.85)}) \Rightarrow |\phi_i^{(\mathrm{new})}\rangle. \tag{5.12}$$

(v) For the updated wave functions $\{\phi_i\}$, the additional Löwdin orthogonalization is imposed:

$$\rho = \sum_i |\phi_i\rangle\langle\phi_i| \tag{5.13}$$

$$|\phi_i\rangle \Rightarrow |\phi_i\rangle - \frac{1}{2}(\rho - 1)|\phi_i\rangle. \tag{5.14}$$

In the program code, the second term of Eq. (5.14) is truncated within the localization region of $|\phi_i\rangle$. This orthogonalization procedure is repeated as an inner loop, until the converge :

$$\{(5.13) \rightarrow (5.14)\} \rightarrow \{(5.13) \rightarrow (5.14)\} \rightarrow \{(5.13) \rightarrow (5.14)\} \rightarrow \cdots$$

Within the above orthogonalization procedure, the quantity

$$\gamma \equiv \sum_i |(\rho - 1)|\phi_i\rangle| \tag{5.15}$$

is monitored. This quantity decreases in course of the above iterative orthogonalization procedure and will be zero, if no localization constraint is imposed. In the program code, the iterative loop is stopped, when the quantity γ stops the exponential decrease. A typical iterative number is two.

The procedures (i)-(v) are shown in Fig. 5.1 as a chart. The above iterative loop is carried out iteratively for the self-consistency.

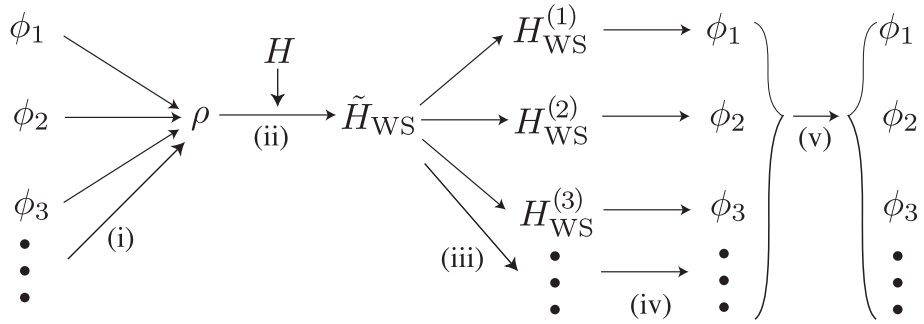


Figure 5.1: Chart of the procedures within one iterative loop for the update of wave functions.

Optimal algorithm in inhomogeneous systems

In some practical simulations, the values of the deviations $|\delta\phi_i|$ are quite different among the Wannier states, due to the inhomogeneous property of the system. A typical example is seen in the brittle fracture simulations (See Chapter 7). The Wannier states in bond breaking processes change their character significantly from the bulk (sp^3) bonding states into surface ones. The other Wannier states keep their character of the bulk (sp^3) bonding states. For an optimal algorithm, the corresponding algorithm should be treated inhomogeneously among the Wannier states. Here we discuss such inhomogeneous treatments among the Wannier states.

As an inhomogeneous treatment, we can prepare the different localization regions among the Wannier states. This treatment will be discussed in Section 7.4. As another inhomogeneous treatment, we explain a choice in the algorithm between the two methods called 'band-by-band' method and 'all-band' method. These names are often used in the standard *ab initio* calculations, when one discusses the similar problem. In the present program code, the wave functions are updated successively. For example, the first Wannier state is updated ($|\phi_1^{(\mathrm{old})}\rangle \rightarrow |\phi_1^{(\mathrm{new})}\rangle$) and then the second Wannier state is updated ($|\phi_2^{(\mathrm{old})}\rangle \rightarrow |\phi_2^{(\mathrm{new})}\rangle$). When the i-th Wannier state is updated, the updated Wannier states for $j = 1, 2, 3...i-1$ ($\{|\phi_j^{(\mathrm{new})}\}_{j=1,2,...i-1}$) are already obtained. The choice is whether the mean-field Hamiltonian for ϕ_i ($H_{\mathrm{WS}}^{(i)}$) should be constructed with the old wave functions ($\{|\phi_j^{(\mathrm{old})}\}_{j=1,2,...i-1}$) or the updated wave functions ($\{|\phi_j^{(\mathrm{new})}\}_{j=1,2,...i-1}$). If we construct the Hamiltonian with the *old* wave functions, we call the method 'all-band method'. If we construct the Hamiltonian with the *updated* wave functions, we call the method 'band-by-band method'. In the band-by-band method, the order among the successive updates is meaningful, while it is meaningless in the all-band method. In our experiences, an optimal choice of the order in the band-by-band method is the order sorted by the values of the deviations $\{|\delta\phi_i|\}$. That is, the Wannier states are sorted so as to satisfy the relation $|\delta\phi_1| > |\delta\phi_2| > |\delta\phi_3| \cdots$. In most cases of practical molecular dynamics, the band-by-band method seems to be better than the all-band method. Among the results of the present thesis, the all-band method is used only in the perfect crystal case, in which all the Wannier states are symmetrically equivalent. Here we add a comment on the band-by-band method; Since the Wannier states are locally determined, two Wannier states that are well separated in real space do not affect with each other. Therefore, a global sorting procedure is not necessary in the above sorting procedure.

Preparation of initial states

Finally, the preparation of the initial Wannier states is discussed. In general, the iteration number for convergence depends on the initial states. During the molecular dynamics simulations, the initial wave functions can be chosen as the final wave functions of the previous time step. In results, a typical iterative number is only one, two or three. Therefore, the main problem is the preparation of the initial states at the *first* time step of the molecular dynamics simulations. In the present simulation of silicon, the initial Wannier states in bulk regions is the perturbative Wannier states based on the sp^3 bonding orbital. For several surface states, a lone pair state on a surface atom is chosen to be the initial Wannier state. These initial states are clearly good candidates, because a bonding state and a lone pair state are typical examples of the Wannier states. For a general case, however, good candidates for the initial Wannier states are not clear. In principle, the Wannier states can be generated rigorously from the eigen states, as explain in Chapter 2.3, using the unitary transformations. Though such a calculation with eigen states is applicable only to small systems, the resultant Wannier states may give an insight for Wannier states in large systems.

5.3 Hybrid scheme

This section is devoted to several technical details in the hybrid scheme introduced in Section 4.4. In this section, the notations and the calculated systems are the same in Section 4.4, except where indicated. The following discussions are done in the hybrid scheme between the diagonalization method for the 'A' subsystem and the perturbative method for the 'B' subsystem.

Comparison with different settings in subspace division

Here we compare the results with different settings in the subspace division. In Section 4.4, the setting of the subspace division was determined by choosing the controlling parameter $z^{(c)}$ and its value was fixed to be $z^{(c)} = 8$. The Hamiltonian matrix $H^{(\mathrm{A})}_{\mathrm{map}}$ was prepared among all the basis set with 1024 atoms, which requires the same computational costs as in the exact diagonalization. As a practical method applicable to large-scale calculations, we prepare the Hamiltonian matrix $H^{(\mathrm{A})}_{\mathrm{map}}$ among a partial basis set. The partial basis set is constructed from the atoms that lie only within $0 \leq z \leq z^{(A)} \equiv 12$. In other words, we ignore the contribution of ρ_A among the region of deeper layers ($z > z^{(A)}$). This approximation is justified, due to the decay property of the charge distribution $n_{\mathrm{A}}(z)$, as shown in Fig. 4.4(b). The choice of $z^{(A)}$ is independent from that of $z^{(c)}$, if they satisfy $z^{(c)} < z^{(A)}$. The distance $(z^{(A)} - z^{(c)})$ corresponds to the cutoff distance for Wannier states. The resultant Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$ is an explicit matrix among $64 \times (z^{(A)}+1) = 64 \times 13 = 832$ atoms, which is slightly smaller than the total number of atoms (1024). The number of the occupied states for the subsystems 'A' and 'B' are unchanged (N_{A} and N_{B}). The results are almost unchanged; For the band of ρ_{A} in Fig. 4.2, the band top and bottom are numerically unchanged. The charge distribution $n_{\mathrm{A}}(z)$ is also almost unchanged from that in Fig. 4.4(a). For example, the value of $n_{\mathrm{A}}(z = 8)$ is unchanged within a numerical error of 10^{-6}.

Now the setting of the subspace division is determined by the two independent controlling parameters $z^{(c)}$ and $z^{(A)}$. Figure 5.2 demonstrates the results with different values of the parameters, among which the parameters $z^{(A)}$ is given as $z^{(A)} = z^{(c)} + 4$. All the results reproduce the charge transfers in the surface region, but quantitative errors are included. When the value of $z^{(c)}$ increases, the size of the 'A' subsystem will increase and the error will decrease. Here we should recall that the present system is unstable due to the unreconstructed surface and is one of the severest test cases for the methodology.

The present discussion was done just for showing the controlling parameters $z^{(c)}$ and $z^{(A)}$ in the present hybrid scheme, which does not make any conclusive remark on how the parameters $z^{(c)}$ and $z^{(A)}$ should be chosen in practical molecular dynamics. Moreover, the present discussion is limited, for simplicity, to a one-dimensional discussion for a slab system. In Chapter 7, we will discuss the hybrid scheme between the variational order-N method and the perturbative order-N method. We will explain that the hybrid scheme contains several controlling parameters and discuss how they are chosen in practical large-scale calculations.

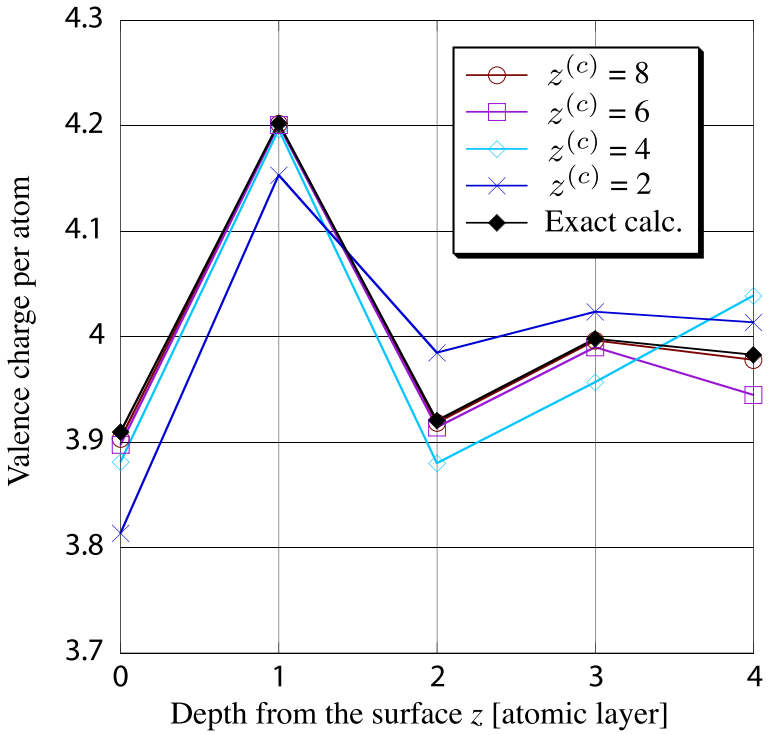


Figure 5.2: The comparison of the results with the different values of the controlling parameter $z^{(c)}$ and $z^{(A)}$. The averaged valence charge per atom with the function of the depth from the surface (z) in the unit of the atomic layer. The system is a silicon crystal with an unreconstructed surface at $z = 0$. The parameters $z^{(A)}$ is given as $z^{(A)} = z^{(c)} + 4$. The result with the exact diagonalization is also plotted.

The choice of the energy shift parameter

Now we discuss the another important controlling parameter in the hybrid scheme, that is, the energy shift parameter $\eta_{\rm s}$. As discussed in Section 4.4, the parameter $\eta_{\rm s}$ should be large enough to separate the band of ρ_B from ρ_A among the eigen value distribution of $H^{\rm (A)}_{\rm map}$, as in Fig. 4.2. The value of $\eta_{\rm s}$ is upper unbounded, if the density matrix ρ_B is constructed from the *exact* Wannier states. In several practical cases, however, the parameter $\eta_{\rm s}$ may be upper bounded. We show such a case with a *too large* value of $\eta_{\rm s}$. Figure 5.3 shows the DOS of the mapped Hamiltonian $H^{\rm (A)}_{\rm map}$, in which the parameter $\eta_{\rm s}$ is chosen as a too large value ($\eta_{\rm s} = 5$a.u.) and all the other controlling parameters are the same as in Fig. 4.2. The band of ρ_B is placed in the energy range of $\varepsilon \approx 2\eta_{\rm s} = 272$ eV and is not plotted in the figure. The result is completely unphysical. For example, several eigen levels of $H^{\rm (A)}_{\rm map}$ are lower than the bottom of the correct valence band ($\varepsilon = -14$ eV). We pick out the five lowest eigen states of $H^{\rm (A)}_{\rm map}$, as examples of the unphysical low energy states. They are unphysical not only in the energy level but also in the character of wave functions. Their weights of s orbitals ($f_{\rm s}^{(i)}$), defined in Eq. (3.9), are quite small ($0.1 \leq f_{\rm s}^{(i)} \leq 0.3$), though the correct wave function at the valence band bottom should be a pure s orbital ($f_{\rm s}^{(i)} = 1$). The highest occupied level for the subsystem ρ_A, determined by Eq. (4.9), is also unphysically low ($\varepsilon = -2.287$ eV), because the

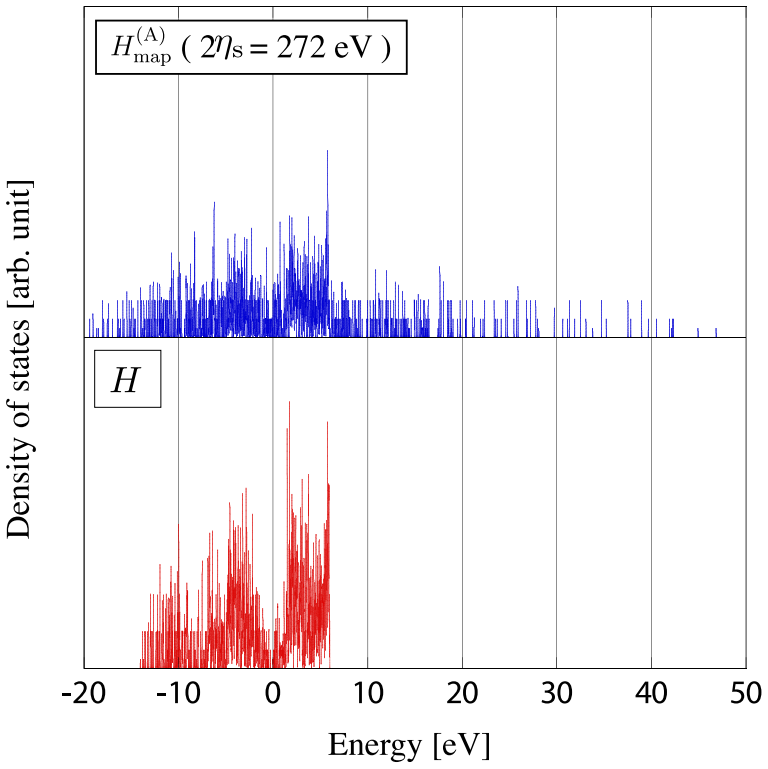


Figure 5.3: The density of states (DOS) of the mapped Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$ with a *too large* value of η_{s} ($\eta_{\mathrm{s}} = 5$a.u.).

above unphysical wave functions are included in the occupied states of ρ_A.

The origin of the unphysical solutions is the finite band width of ρ_B at the domain $x \approx 0$ in Fig. 4.3. The finite band width is estimated as $\Delta x = 0.2$. The mapped Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$ contains the term of $2\eta_{\mathrm{s}}\rho_B$ and the corresponding band width is estimated as $2\eta_{\mathrm{s}} \times \Delta x = 2$a.u. ≈ 54.4eV, which is comparable to the band width in the observed DOS in Fig. 5.3. Figure 5.4 shows the density distribution of the unphysical wave functions. They are extended states in the region of $z \geq 6$, a region of deeper layers. This property can be understood as follows; The finite bandwidth in the DOS profile of ρ_B originates from the error of the orthogonality among the perturbative Wannier states ($\langle \phi^{(\mathrm{PT})}_i | \phi^{(\mathrm{PT})}_j \rangle \neq \delta_{ij}$) and such Wannier states lie in the region of $z \geq 6$, as discussed above. The off-diagonal elements of the overlap matrix $\langle \phi^{(\mathrm{PT})}_i | \phi^{(\mathrm{PT})}_j \rangle$ works formally as a transfer matrix in the Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$. With a large value of η_{s}, the contribution of $2\eta_{\mathrm{s}}\rho_B$ is dominant in the Hamiltonian $H^{(\mathrm{A})}_{\mathrm{map}}$, and its lowest eigen state should be an extended state among the region of $z \geq 6$, so as to gain the transfer energy due to the non-zero overlap matrix.

Fortunately, the above unphysical solutions are automatically excluded, when the 'A' subsystem is determined by the variational order-N method with an explicit localization constraint on Wannier states. For the Wannier states in the 'A' subsystem, their localization centers are placed in the region near the surface ($z < z^{(c)}$) and several localization constraints are imposed with a cutoff radius. The explicit localization constraint on Wannier states is equivalent to the situation with the additional potential wall in the infinite hight ($V(\boldsymbol{r}) = \infty$). Such localization constraints

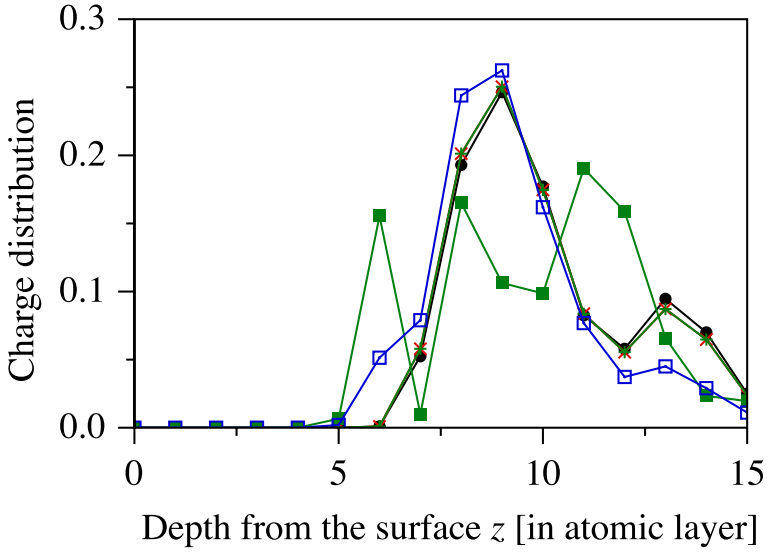


Figure 5.4: The charge distribution of the lowest five eigen states of $H^{(\mathrm{A})}_{\mathrm{map}}$ with a *too large* value of η_{s} ($\eta_{\mathrm{s}} = 5$a.u.). The system is a silicon crystal with an unreconstructed surface at $z = 0$.

prohibit the solutions from the unphysical extended wave functions in Fig. 5.4. In other words, the localization constraint is an additional mechanism for the exclusion of the 'B' subsystem from the 'A' subsystem.

5.4 Parallelization of perturbative order-N method

Now the parallelization algorithm is discussed within the perturbative order-N method. Since the procedures among Wannier states or among the 'domains' are completely independent, as explained in Section 4.3 and 5.1, their parallelization is straightforward, at least, logically.

The Message Passing Interface (MPI) technique [83] and the OpenMP technique [84] are the standard techniques in present parallel computations. We implement, experimentally, both techniques and test them using a workstation cluster (SGI Origin 3800). Here the result with the OpenMP technique is discussed. The result with the MPI technique is discussed in Ref. [80]. The computational costs depend on various factors in parallel computers, such as the data communication between processors. We have not yet settled the details of the program code, which may affect on the resultant elapse time by a factor, say, two or three. Figure 5.5 shows a result using the OpenMP technique with up to $N_{\rm P} = 256$ processors. The sample is a silicon cluster with about 1.4 million atoms. The solid line indicates the elapse time for 'electronic' part, that is, the procedures of the perturbative order-N method. The dashed line indicates the ideal parallel efficiency that is inversely proportional to the number of processors ($N_{\rm P}$). The dashed line is plotted so as to cross the result with $N_{\rm P} = 4$. Note that the result with the single processor ($N_{\rm P} = 1$) is obtained by the non-parallelized code, in which no parallel directive is included. The resultant elapse time at $N_{\rm P} = 1$ is slightly, by about 15 %, deviated from the ideally parallel (dashed) line. The elapse time for 'other part' indicates the elapse time without the above 'electronic' part. This part is equivalent, in the computational costs, to a program code of a classical molecular dynamics with a two-body short-range potential. The latter part is implemented in an order-N computational cost but have not yet been parallelized.

From the elapse times in the single processor ($N_{\rm P} = 1$), we can find that the elapse time of the 'electronic' part is more than 200 times larger than that of the 'other' part. This ratio is based on the quantum mechanical prefactor $\nu^2 = 16$ discussed in Section 2.4. Since the value of 16 is the minimal prefactor, the observed prefactor of 200 is a reasonable value for the practical code. Due to the above large prefactor, we do not need to parallelize the 'other' part, the routines for classical molecular dynamics, at least, using less than 100 processors.

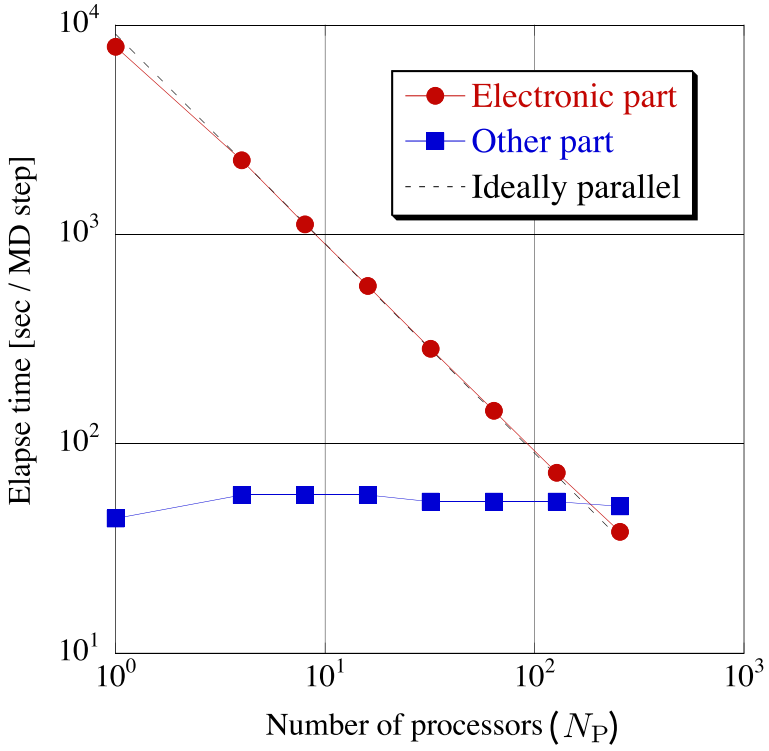


Figure 5.5: The elapse time of the perturbative order-N calculations in the parallel computations The OpenMP technique is used with up to $N_{\rm P} = 256$ processors. The system is a silicon cluster with 1,423,909 atoms. We measure the elapse time per one time step in the molecular dynamics simulation. The 'electronic' part indicates the procedures with the wave functions. The 'other' part indicates the procedures that are not related to the wave functions. The 'other' part is not parallelized.

5.5 Wannier states in diamond structure solids

In this section, the Wannier states in the diamond structure solids are systematically investigated [27, 85]. The present investigation is based on the universal tight-binding theory (See Section 3.1). Particularly, we focus on the following points; (i) The structure of the Wannier state, especially the approximate wave function using the perturbative formulation (See Section 4.3). (ii) How the density matrix or the electronic structure energy is reproduced, quantitatively, by approximate Wannier states. (iii) The difference of the Wannier states among the group IV elements.

Among the above points, we will see the crucial importance of the mixing freedom between the s and p bands. In general, the mixing or hybridization freedom between bands is characteristic to the generalized or composite-band Wannier states (See Section 2.3). In Appendix D.1, on the other hand, a conventional or isolated-band Wannier state is constructed using the unitary transforms of eigen states *without mixing bands*.

Hamiltonian matrix elements with sp^3 orbitals

First, the nearest neighbor tight-binding Hamiltonian is described as the explicit matrix with sp^3 hybridized atomic orbitals. Within an atom pair of the diamond structure, the eight sp^3 hybridized atomic orbitals are defined, which lie on the bond sites ($\{|\mathrm{h}i\rangle\}, i = 1, 2, \cdots 8$). The geometry of the eight orbitals are shown in Fig. 5.6. Figure 5.6(a) indicates the definition of the orbitals with respect to the sign freedom. Hereafter the above sp^3 orbitals are used as the basis set of the tight-binding Hamiltonian matrix H. The diagonal element is denoted as ε_{h};

$$\varepsilon_{\mathrm{h}} \equiv \frac{\varepsilon_{\mathrm{s}} + 3\varepsilon_{\mathrm{p}}}{4}. \tag{5.16}$$

The off-diagonal elements within the same atom, or the intraatomic hoppings, have the unique value of

$$\beta_0 \equiv -\frac{\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}}}{4}. \tag{5.17}$$

The interatomic hoppings, on the other hand, are classified into four values;

$$\beta_1 \equiv \frac{1}{4}\left\{V_{\mathrm{ss}\sigma} - 2\sqrt{3}V_{\mathrm{sp}\sigma} - 3V_{\mathrm{pp}\sigma}\right\} \tag{5.18}$$

$$\beta_2 \equiv \frac{1}{4}\left\{V_{\mathrm{ss}\sigma} - \frac{2}{\sqrt{3}}V_{\mathrm{sp}\sigma} + V_{\mathrm{pp}\sigma}\right\} \tag{5.19}$$

$$\beta_3 \equiv \frac{1}{4}\left\{V_{\mathrm{ss}\sigma} + \frac{2}{\sqrt{3}}V_{\mathrm{sp}\sigma} - \frac{1}{3}V_{\mathrm{pp}\sigma} - \frac{8}{3}V_{\mathrm{pp}\pi}\right\} \tag{5.20}$$

$$\beta_4 \equiv \frac{1}{4}\left\{V_{\mathrm{ss}\sigma} + \frac{2}{\sqrt{3}}V_{\mathrm{sp}\sigma} - \frac{1}{3}V_{\mathrm{pp}\sigma} + \frac{4}{3}V_{\mathrm{pp}\pi}\right\}. \tag{5.21}$$

The values are shown in Table 5.1. Among the interatomic hoppings, the dominant one is β_1, that is, the hopping along the bond. For example, $\langle \mathrm{h1}|H|\mathrm{h5}\rangle = \beta_1$ in Fig. 5.6.

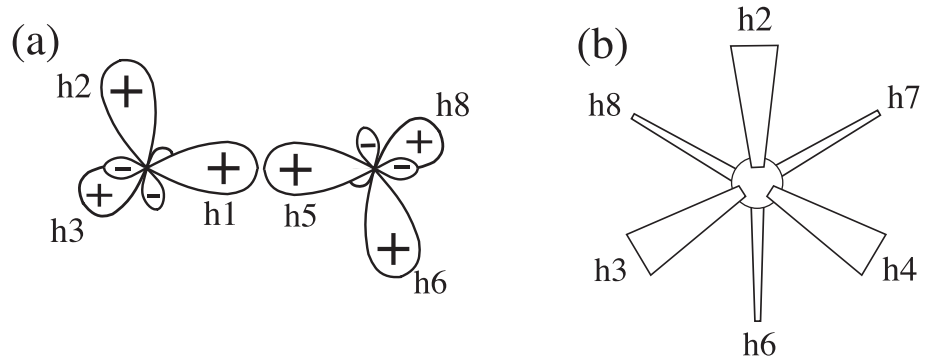


Figure 5.6: The geometry of the eight sp^3 hybridized atomic orbitals ($\{|hi\rangle\}, i = 1, 2, \cdots 8$), in the diamond structure. In (a), the four orbitals $|h2\rangle, |h1\rangle, |h5\rangle, |h6\rangle$, forms a part of the zigzag chain within the (110) plane. The sign freedom of sp^3 orbitals are shown in (a), which is essential to determine the orbital uniquely.

	h1	h2	h3	h4
h1	ε_h	β_0	β_0	β_0
h2	β_0	ε_h	β_0	β_0
h3	β_0	β_0	ε_h	β_0
h4	β_0	β_0	β_0	ε_h

	h5	h6	h7	h8
h1	β_1	β_2	β_2	β_2
h2	β_2	β_3	β_4	β_4
h3	β_2	β_4	β_3	β_4
h4	β_2	β_4	β_4	β_3

	ε_s	$\varepsilon_p - \varepsilon_s$	ε_h	β_0	β_1	β_2	β_3	β_4
C	-2.99	6.70	2.04	-1.68	-9.45	-1.23	0.68	-0.87
Si	-5.25	6.45	-0.41	-1.61	-4.08	-0.33	0.48	-0.59

Table 5.1: Upper tables: the classification of the non-zero elements in the minimal tight-binding Hamiltonian with the basis set of sp^3 orbitals. The geometry of each sp^3 orbitals are shown in Fig. 5.6. Lower table: the value of the matrix elements in the carbon [64] and silicon [6] case in the energy unit of eV.

As already discussed in Section 3.1, a simple bonding and antibonding orbitals are defined as

$$\begin{aligned} |b\rangle &= \frac{|hi\rangle + |hj\rangle}{\sqrt{2}} \\ |a\rangle &= \frac{|hi\rangle - |hj\rangle}{\sqrt{2}}, \end{aligned} \tag{5.22}$$

with a pair of the sp^3 orbitals on a bond site ($|hi\rangle, |hj\rangle$). The corresponding energy levels are obtained by

$$\begin{aligned} \varepsilon_b &\equiv \langle b|H|b\rangle = \varepsilon_h + \beta_1 \\ \varepsilon_a &\equiv \langle a|H|a\rangle = \varepsilon_h - \beta_1, \end{aligned} \tag{5.23}$$

where the value β_1 is negative ($\beta_1 < 0$).

Wannier state in silicon

As explained in Section 4.3, the first-order perturbation of Eq.(4.94) gives an approximate Wannier state. Here we rewrite the formulation

$$|\phi_i^{(\mathrm{PT})}\rangle = C^{(0)}|\mathrm{b}_i\rangle + \sum_{j(\neq i)} C^{(\nu(j))}|\mathrm{a}_j\rangle \tag{5.24}$$

Here the suffix i indicates a bond site in diamond structure, which is the localization center of the i-th Wannier state. The suffix ν specifies the inequivalent bond sites. The coefficients will be given explicitly below.

The central region of the Wannier state is shown in Fig. 5.7, which includes up to the second nearest neighbor bond sites. The inequivalent bond sites are indicated as $\nu = 1, (2\,\|)\,\mathrm{or}\,(2\,\perp)$. The central bond site is marked as (0). The six first nearest neighbor bond sites are symmetrically equivalent and are marked as (1). The eighteen second nearest neighbor bond sites are classified into two symmetrically inequivalent sites, which are marked as $(2\,\|)$ and $(2\,\perp)$. Among them, the six bonds site, marked $(2\,\|)$, are parallel to the central bond. The other twelve bond sites, marked $(2\,\perp)$, are nearly perpendicular to the central bond. Now the (anti)bonding orbitals should be uniquely defined with respect to the sign freedom. The definition is given in Fig. 5.8, which determines the sign of the perturbative coefficients $C^{(\nu(j))}$ uniquely.

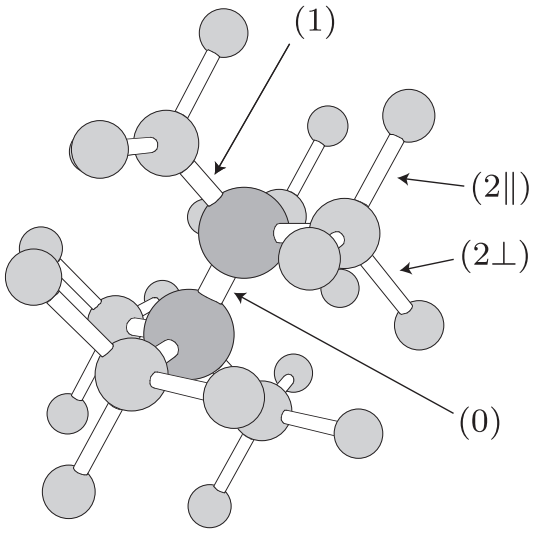


Figure 5.7: The central region of the Wannier state in the diamond structure [85]. The size of ball distinguishes the inequivalent atom sites. The central bond site is marked as (0) and one of the first nearest-neighbor bond site is marked as (1). Some of the two inequivalent second nearest-neighbor bond sites are marked as $(2\,\|)$ and $(2\,\perp)$.

In the perturbative formulation, the spatial spread of the Wannier states is determined by the hopping range of the Hamiltonian. In the case of a Hamiltonian with

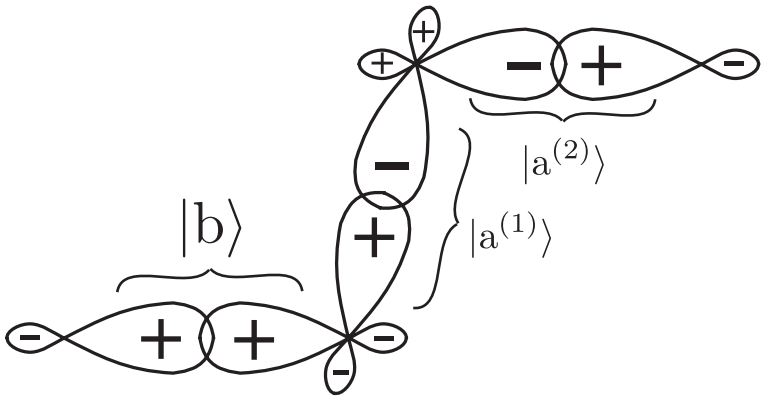


Figure 5.8: Definition of (anti)bonding orbitals with respect to the sign freedom of wave function, which is essential to determine uniquely the values of the coefficients in Eq. 5.25 and Eq. 5.26. The central bonding orbital and the first and second nearest neighbor antibonding orbitals are denoted as $|\mathrm{b}\rangle$, $|\mathrm{a}^{(1)}\rangle$ and $|\mathrm{a}^{(2)}\rangle$, respectively.

only the nearest neighbor hoppings between atoms, it turns to be the *second* nearest neighbor hoppings between bond sites. For the first nearest neighbor antibonding orbitals, the perturbative coefficients are given [62] by

$$\frac{C^{(1)}}{C^{(0)}} \approx \frac{\langle a^{(1)}|H|b_k\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}} = \frac{\beta_0/2}{2\beta_1} = \frac{(\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}})/8}{2\beta_1} = \frac{\alpha_{\mathrm{m}}}{8}. \tag{5.25}$$

For the second nearest neighbor anti-bonding orbitals, we propose the following coefficients

$$\frac{C^{(2\lambda)}}{C^{(0)}} \approx \frac{\langle a^{(2\lambda)}|H|b_i\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}} + \left(\frac{\alpha_{\mathrm{m}}}{8}\right)^2, \tag{5.26}$$

where (2λ) indicates $(2\parallel)$ or $(2\perp)$. The first term is the direct hopping term. The second term is responsible for the successive hopping of the first nearest neighbor hoppings, where $C^{(0)} = 1$ is assumed. Note that the second term does not appear in the standard perturbation theory of Eq. (4.95). The value of the first term is different among the two inequivalent bond sites;

$$\frac{\langle a^{(2\parallel)}|H|b_i\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}} = \frac{\beta_3/2}{2\beta_1} \approx \frac{1}{34} \tag{5.27}$$

or

$$\frac{\langle a^{(2\perp)}|H|b_i\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}} = \frac{\beta_4/2}{2\beta_1} \approx -\frac{1}{28}. \tag{5.28}$$

Here the opposite signs between β_3 and β_4 results in the opposite signs between the above coefficients. The resultant values of the perturbation coefficients are given as

$$\begin{aligned} \frac{C^{(1)}}{C^{(0)}} &= \frac{\alpha_{\mathrm{m}}}{8} = \frac{0.78}{8} = 0.0975 \\ \frac{C^{(2\parallel)}}{C^{(0)}} &= \frac{\langle a^{(2\parallel)}|H|b_i\rangle}{\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{a}}} + \left(\frac{\alpha_{\mathrm{m}}}{8}\right)^2 \end{aligned} \tag{5.29}$$

$$\approx \frac{1}{34} + (0.0975)^2 \approx 0.0389 \tag{5.30}$$

$$\frac{C^{(2\perp)}}{C^{(0)}} = \frac{\langle a^{(2\perp)}|H|b_i\rangle}{\varepsilon_{\rm b} - \varepsilon_{\rm a}} + \left(\frac{\alpha_{\rm m}}{8}\right)^2$$

$$\approx -\frac{1}{28} + (0.0975)^2 \approx -0.0262. \tag{5.31}$$

The total weight within the second nearest neighbor bond sites is defined as

$$\mathcal{N}_2 \equiv |C^{(0)}|^2 + 6|C^{(1)}|^2 + 6|C^{(2\parallel)}|^2 + 12|C^{(2\perp)}|^2. \tag{5.32}$$

The normalization condition

$$\mathcal{N}_2 = 1 \tag{5.33}$$

is imposed so as to determine the value of $C^{(0)}$. In results, the Wannier state $|\phi_i^{(\rm PT)}\rangle$ in Eq. (5.24) is determined uniquely by Eqs. (5.25), (5.26), (5.33).

The resultant values of the coefficients are shown in Table 5.2 with the corresponding values of the exact Wannier state. Here the exact Wannier state is calculated in the periodic cell of 512 atoms, without any localization constraint. The resultant one-electron energy $\langle\phi_i|H|\phi_i\rangle$ has an error of 0.054 eV from the exact value. Other physical quantities have been already discussed in Section 4.3.

	$\lvert C^{(0)}\rvert^2$	$\lvert C^{(1)}/C^{(0)}\rvert^2$	$\lvert C^{(2\parallel)}/C^{(0)}\rvert^2$	$\lvert C^{(2\perp)}/C^{(0)}\rvert^2$	$\mathcal{N}_2$
Perturbative	0.934	0.00904	0.00151	0.000686	1
Exact	0.938	0.00670	0.00250	0.000499	0.995

Table 5.2: Values of the perturbative coefficients in silicon, given by Eqs. (5.25), (5.26), (5.33) and the corresponding values that is by the exact calculation with 512 atoms. The total weight within the second nearest neighbor bond sites $\mathcal{N}_2$ is also shown.

Density matrix

Now we demonstrate how the density matrix is constructed, quantitatively, from the Wannier states. For simplicity, we consider only two successive bond sites, as in Fig. 5.9, with the four sp^3 atomic orbitals, $|{\rm h_I}\rangle, |{\rm h_{II}}\rangle, |{\rm h_{III}}\rangle, |{\rm h_{IV}}\rangle$.

Two approximate Wannier states are given by

$$|\phi_1\rangle \approx \frac{1}{\sqrt{2}}(|{\rm h_I}\rangle + |{\rm h_{II}}\rangle) + \frac{C^{(1)}}{\sqrt{2}}(|{\rm h_{III}}\rangle - |{\rm h_{IV}}\rangle)$$
$$+(\text{terms on other basis}) \tag{5.34}$$

$$|\phi_2\rangle \approx \frac{1}{\sqrt{2}}(|{\rm h_{III}}\rangle + |{\rm h_{IV}}\rangle) + \frac{C^{(1)}}{\sqrt{2}}(|{\rm h_{II}}\rangle - |{\rm h_I}\rangle)$$
$$+(\text{terms on other basis}), \tag{5.35}$$

whose centers are located on the two bond sites. These forms are simpler ones from Eq. (5.24) in the sense that we ignore the normalization factor ($C^{(0)}$) and

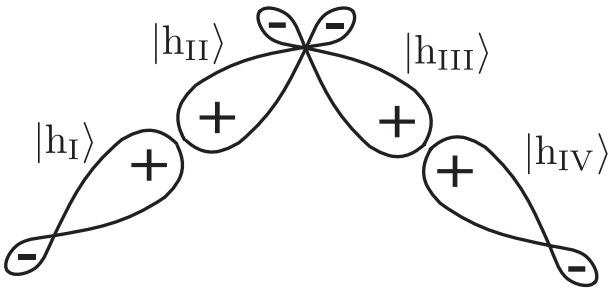


Figure 5.9: Schematic pictures of the sp^3 atomic orbitals $|h_I\rangle$, $|h_{II}\rangle$, $|h_{III}\rangle$, $|h_{IV}\rangle$.

	$\langle h_i\|H\|h_j\rangle$ (eV)	$\langle h_i\|\rho\|h_j\rangle_{\rm exact}$	$\langle h_i\|\rho\|h_j\rangle_{\rm est}$
$(i,j)=(I,II)$	-4.08 $(=\beta_1)$	0.439	0.45
$(i,j)=(II,III)$	-1.61 $(=\beta_0)$	0.078	0.09
$(i,j)=(I,III)$	-0.33 $(=\beta_2)$	-0.008	0
$(i,j)=(I,IV)$	0	-0.071	-0.09

Table 5.3: Elements of the Hamiltonian matrix and the density matrix on sp^3 orbitals bases shown in Fig. 5.9. The values of $\langle h_i|\rho|h_j\rangle_{\rm exact}$ are the results of the diagonalization with 512 atoms. The values of $\langle h_i|\rho|h_j\rangle_{\rm est}$ are the estimated one in Eqs. (5.37)-(5.40) with a normalization coefficient of $|C^{(0)}|^2 \approx 0.9$.

the contributions of the second nearest neighbor bond sites ($C^{(2)}$). The one-body density matrix is given by

$$\rho \approx |\phi_1\rangle\langle\phi_1| + |\phi_2\rangle\langle\phi_2| \tag{5.36}$$

on the present four sp^3 orbitals. The other Wannier states ($\{\phi_i\}_{i\neq 1,2}$) do not contribute the density matrix on the present orbitals, within the linear order of the perturbation coefficient ($C^{(1)}$). The matrix elements on the sp^3 orbitals are calculated as

$$\begin{aligned}
\langle h_I|\rho|h_{II}\rangle &= \langle h_I|\phi_1\rangle\langle\phi_1|h_{II}\rangle + \langle h_I|\phi_2\rangle\langle\phi_2|h_{II}\rangle \\
&= \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}} + \frac{-C^{(1)}}{\sqrt{2}}\frac{C^{(1)}}{\sqrt{2}} \approx 0.5 \qquad (5.37)\\
\langle h_{II}|\rho|h_{III}\rangle &= \langle h_{II}|\phi_1\rangle\langle\phi_1|h_{III}\rangle + \langle h_{II}|\phi_2\rangle\langle\phi_2|h_{III}\rangle \\
&= \frac{1}{\sqrt{2}}\frac{C^{(1)}}{\sqrt{2}} + \frac{C^{(1)}}{\sqrt{2}}\frac{1}{\sqrt{2}} \approx 0.1 \qquad (5.38)\\
\langle h_I|\rho|h_{III}\rangle &= \langle h_I|\phi_1\rangle\langle\phi_1|h_{III}\rangle + \langle h_I|\phi_2\rangle\langle\phi_2|h_{III}\rangle \\
&= \frac{1}{\sqrt{2}}\frac{C^{(1)}}{\sqrt{2}} + \frac{-C^{(1)}}{\sqrt{2}}\frac{1}{\sqrt{2}} \approx 0 \qquad (5.39)\\
\langle h_I|\rho|h_{IV}\rangle &= \langle h_I|\phi_1\rangle\langle\phi_1|h_{IV}\rangle + \langle h_I|\phi_2\rangle\langle\phi_2|h_{IV}\rangle \\
&= \frac{1}{\sqrt{2}}\frac{-C^{(1)}}{\sqrt{2}} + \frac{-C^{(1)}}{\sqrt{2}}\frac{1}{\sqrt{2}} \approx -0.1 \qquad (5.40)
\end{aligned}$$

within the linear order of $C^{(1)} \approx 0.1$. Table 5.3 shows the estimated values of the density matrix $\langle h_i|\rho|h_j\rangle_{\rm est}$ that are determined by the above expressions with

multiplying the normalization factor of $|C^{(0)}|^2 \approx 0.9$. The table also shows the corresponding exact values of the density matrix $\langle \mathrm{h}_i|\rho|\mathrm{h}_j\rangle_{\mathrm{exact}}$ and those of the Hamiltonian matrix $\langle \mathrm{h}_i|H|\mathrm{h}_j\rangle$. Though the above estimation is quite simple, the resultant values reproduce satisfactory the exact values.

Among the resultant density matrix elements, the comparison of $\langle \mathrm{h_I}|\rho|\mathrm{h_{III}}\rangle$ and $\langle \mathrm{h_I}|\rho|\mathrm{h_{IV}}\rangle$ is interesting. In the former case, the matrix element is quite small ($\langle \mathrm{h_I}|\rho|\mathrm{h_{III}}\rangle = -0.007$), though the corresponding hopping integral is finite (β_2). In the latter case, the density matrix is non negligible ($\langle \mathrm{h_I}|\rho|\mathrm{h_{IV}}\rangle = -0.069$), though the corresponding hopping integral is zero. The above facts are explained by the quantum mechanical interference of the two Wannier states, as in Eqs. (5.39) and (5.40).

Energy

Due to the symmetry in diamond structure, the one-electron energy of the Wannier state $\varepsilon_{\mathrm{WS}} = \langle\phi_i|H|\phi_i\rangle$ gives the average of the occupied eigen levels (Eq. (4.73)). Here the energy $\varepsilon_{\mathrm{WS}}$ is estimated by the perturbative formulation without the normalization condition and the second nearest neighbor terms;

$$\varepsilon_{\mathrm{WS}} \approx \varepsilon_{\mathrm{b}} - 6\frac{(\beta_0/2)^2}{\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}}. \tag{5.41}$$

The factor six is the number of the first nearest neighbor bond sites. The second term can be written by

$$\begin{aligned} 6\frac{(\beta_0/2)^2}{\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}} &= 6\left(\frac{\beta_0/2}{\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}}\right)^2 \times (\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}) \\ &= 6\left(\frac{\alpha_{\mathrm{m}}}{8}\right)^2 \times 2\,|\beta_1| \\ &= \frac{3}{16}\alpha_{\mathrm{m}}^2\,|\beta_1|. \end{aligned} \tag{5.42}$$

Using $\alpha_{\mathrm{m}} = 0.78$ and $|\beta_1| = 4.08$ eV, the numerical result is obtained as

$$\begin{aligned} \frac{3}{16}\alpha_{\mathrm{m}}^2\,|\beta_1| &= \frac{3}{16}(0.78)^2 \times 4.08\,[\mathrm{eV}] \\ &= 0.114 \times 4.08\,[\mathrm{eV}] \\ &= 0.465\,[\mathrm{eV}], \end{aligned} \tag{5.43}$$

which explains about 80 % of the exact value of $\varepsilon_{\mathrm{WS}} - \varepsilon_{\mathrm{b}} \approx 0.59$ eV. Since the energy ε_{b} is that of an ideal sp^3 bonding orbital on a bond site, the energy difference $\varepsilon_{\mathrm{WS}} - \varepsilon_{\mathrm{b}}$ corresponds to the energy gain of the Wannier state for its spatial extension in condensed matters.

As an overview of the cohesive mechanism, Fig. 5.10 shows several energy levels of silicon. From Fig. 5.10, we find that the energy gain for the sp^3 bonding ($\varepsilon_{\mathrm{h}} - \varepsilon_{\mathrm{b}}$) is much larger than the energy gain for its spatial extension ($\varepsilon_{\mathrm{b}} - \varepsilon_{\mathrm{WS}}$). We also find that the sp^3 bonding is governed by the energy competition between the bonding energy, scaled by $\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}$ and the dehybridization energy, scaled by $\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}}$, as is discussed in Section 3.1. This competitive situation is characterized by the metallicity parameter $\alpha_{\mathrm{m}} \equiv (\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{s}})/(\varepsilon_{\mathrm{a}} - \varepsilon_{\mathrm{b}}) = 0.78$.

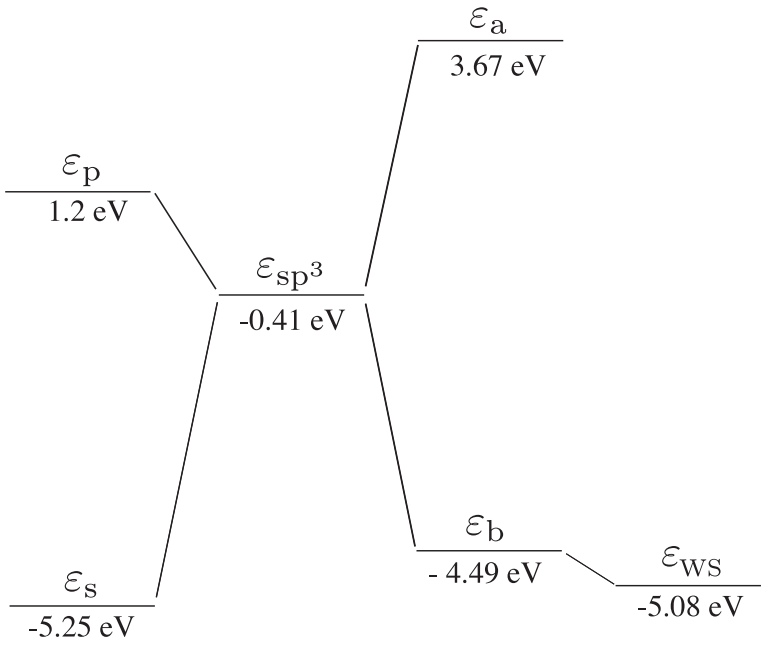


Figure 5.10: Several energy levels of the silicon crystal within the tight-binding Hamiltonian; The atomic s, p and sp^3 levels are denoted as $\varepsilon_{\rm s}$, $\varepsilon_{\rm p}$ and $\varepsilon_{\rm sp^3}$, respectively. The energy levels of the ideal sp^3 hybridized bonding and antibonding orbitals are denoted as $\varepsilon_{\rm b}$ and $\varepsilon_{\rm a}$, respectively. The energy level of the Wannier states is denoted as $\varepsilon_{\rm WS}$, which is the weighted center of the occupied (valence) band.

Wannier states in conduction band

Here we introduce an interesting concept, that is, the Wannier state in the *conduction* band. We discussed that the Wannier state in the valence band is quite similar to an sp^3-hybridized bonding orbital, as in Eq. (5.24). Here we propose an wave function for the Wannier state in the conduction band

$$|\phi^{\rm (PT)}_{i({\rm cond})}\rangle \equiv C^{(0)}|{\rm a}_i\rangle + \sum_{j(\neq i)} C^{(\nu(j))}|{\rm b}_j\rangle, \tag{5.44}$$

where the coefficients are the same values as in Eq. (5.24). The resultant wave function is orthogonal to the corresponding Wannier state in the valence band

$$\langle\phi^{\rm (PT)}_{i({\rm cond})}|\phi^{\rm (PT)}_{i}\rangle = 0. \tag{5.45}$$

Between Eq. (5.24) and Eq. (5.44), the role of the bonding and antibonding orbitals ($\{|{\rm b}_i\rangle, |{\rm a}_i\rangle\}$) exchange with each other. One may think that the relation of $|\phi^{\rm (PT)}_{i({\rm cond})}\rangle$ and $|\phi^{\rm (PT)}_{i}\rangle$ is analogous to the electron-hole symmetry. It should be noted, however, that the present Hamiltonian H does not have the electron-hole symmetry.

The above concept of the Wannier states in the conduction band can be also defined by the variational method, because the present tight-binding Hamiltonian is upper bounded. The Wannier states in the conduction band can be obtained by the energy *maximization* procedure of the energy functional that is used for the

Wannier state in the valence band. As the iterative procedure, the initial state is chosen as the antibonding orbitals $\{|\mathrm{a}_i\rangle\}_{i=1,N}$ and the energy shift parameter η_s is chosen to be sufficiently *low* ($\eta_\mathrm{s} \to -\infty$). The resultant Wannier states satisfy Eq.(4.47), if the N Wannier states are interpreted as those in the conduction band. It is noteworthy that the Wannier states in the valence and conduction bands form a complete orthogonal basis set for the present Hamiltonian.

Construction of Wannier states by projection methods

We also introduce several methods to construct Wannier states, which are different from the variational or perturbative procedure. They were used for the calculation of the *exact* Wannier state within small systems [85]. The sample contains 512 atoms and no localization constraint is imposed on Wannier states. To construct the exact Wannier states, a projection method is used with a 'reference' density matrix ρ_ref.

The practical procedure is as follows; (i) Diagonalizing the Hamiltonian H to obtain the eigen states $\{\phi_k^{(\mathrm{eig})}\}$. The exact ground-state density matrix is calculated as the reference density matrix ρ_ref

$$\hat{\rho}_\mathrm{ref} \equiv \hat{\rho}_\mathrm{GS} \equiv \sum_k^{\mathrm{occ.}} |\phi_k^{(\mathrm{eig})}\rangle\langle\phi_k^{(\mathrm{eig})}|. \tag{5.46}$$

(ii) Preparing proper initial states $\{\phi_j^{(0)}\}$ for Wannier states which covers the set of the correct ground state wave functions. In the present cases of diamond structure solids, we use a set of 'simple' bonding orbitals $\{\mathrm{b}_j\}$ on each bond site

$$|\phi_j^{(0)}\rangle \equiv |\mathrm{b}_j\rangle. \tag{5.47}$$

(iii) Projecting the 'reference' density matrix $\hat{\rho}_\mathrm{GS}$ on the initial states

$$|\phi_j^{(0)}\rangle \Rightarrow |\phi_j^{(1)}\rangle \equiv \hat{\rho}_\mathrm{ref}|\phi_j^{(0)}\rangle. \tag{5.48}$$

(iv) Using the Löwdin orthogonalization, Eq. (5.14), iteratively.

The resultant Wannier states satisfy the mean-field equation (4.62) exactly, which corresponds to the variational order-N method *without any localization constraint*. Note that this method was used for construction of initial states in an iterative order-N procedure with the approximate density matrix [86]. Wannier states *in the conduction band* can be also constructed by a similar projection method. As the 'reference' density matrix, we use that in the conduction band

$$\hat{\rho}_\mathrm{ref} \equiv \hat{\rho}_\mathrm{c} \equiv \sum_k^{\mathrm{unocc.}} |\phi_k^{(\mathrm{eig})}\rangle\langle\phi_k^{(\mathrm{eig})}|, \tag{5.49}$$

instead of ρ_GS in Eq. (5.46). In the iterative procedure, *antibonding* orbitals $\{\mathrm{a}_j\}$ are prepared as the initial states $\{\phi_j^{(0)}\}$.

Moreover, it may be also interesting to apply the above procedure formally to metallic systems. Instead of the ground state density matrix, we use a finite-temperature formulation

$$\hat{\rho}_\mathrm{ref} \equiv \hat{\rho}_\tau \equiv \sum_{k=1}^{\infty} f_\tau(\varepsilon_k)|\phi_k^{(\mathrm{eig})}\rangle\langle\phi_k^{(\mathrm{eig})}| \tag{5.50}$$

where $f_\tau(\varepsilon)$ is a Fermi-Dirac function with a temperature parameter τ. The resultant one-electron states satisfy the correct orthogonality and give a generalized feature of the present Wannier states, though the sum of the one-electron energy of the Wannier states

$$\sum_{j}^{\text{occ.}} \langle \phi_j | \hat{H} | \phi_j \rangle \tag{5.51}$$

is no more equal to the energy of the reference density matrix

$$\text{Tr}[\hat{\rho}_{\text{ref}} \hat{H}] = \sum_{k=1}^{\infty} f_T(\varepsilon_k^{(\text{eig})}) \varepsilon_k. \tag{5.52}$$

The Wannier states constructed from the above projection methods will appear later in this section.

Wannier states among different elements (1)

So far, in this section, we focused on the silicon case. Hereafter, we will investigate systematically the Wannier states among the diamond structure solids (C, Si, Ge and α-Sn). Due to the universal tight-binding theory, the Wannier states among the above elements can be described by the difference of the metallicity parameter α_{m}. Typical values of the metallicity parameter α_{m} are $\alpha_{\text{m}} = 0.35$ for C and $\alpha_{\text{m}} = 0.75 - 0.78$ for Si or Ge (See Section 3.2). For a systematic investigation, we tune the bond length d in the Hamiltonian for silicon [6], while the atomic energy difference $(\varepsilon_{\text{p}} - \varepsilon_{\text{s}})$ is fixed. The above tuning corresponds to the tuning of the metallicity α_{m}. By definition, the silicon case corresponds to the case with $d = d_0 \equiv 2.35$ Å and $\alpha_{\text{m}} = 0.78$. The case with $d = 0.8\, d_0$ gives the value of $\alpha_{\text{m}} = 0.47$, which will be referred as the carbon case in the sense of a low metallicity case.

Figure 5.11 demonstrates the spatial spread of the Wannier state calculated by the variational or perturbative order-N method [27]. In the variational method, the periodic cubic cell with $N_{\text{A}} = 4096$ atoms is used and each Wannier state contains about 150 atoms in its localization constraint. In the perturbative method, the wave functions are analytically determined by Eqs. (5.25),(5.26),(5.33), which is independent from the simulation cell. The case (a) is the carbon case ($d = 0.8 d_0$) and the case (b) is the silicon case ($d = d_0$). The Wannier state is the conduction band is also plotted in Fig.5.11(c) in the silicon case ($d = d_0$). The wave function is constructed in the variational procedure by the energy *maximization* procedure, as explained above. The resultant wave function shows the similar decay property as in the valence Wannier state (b), and the role of bonding and antibonding orbitals seems to exchange with each other approximately. The Wannier state is the conduction band is constructed also from the perturbative method using Eq. (5.44). In all the cases, the variational method results in well-localized wave functions. The weight of the central bond is about 96 % in (a) or 94 % in (b) and (c). The summation of the weight up to the bond step of $n = 2$ is more than 99.7 % in all the cases.

Figure 5.12 also shows the spatial spread of the calculated Wannier state [85]. In Fig. 5.12, the closed circle and open square are plotted by the same plotting

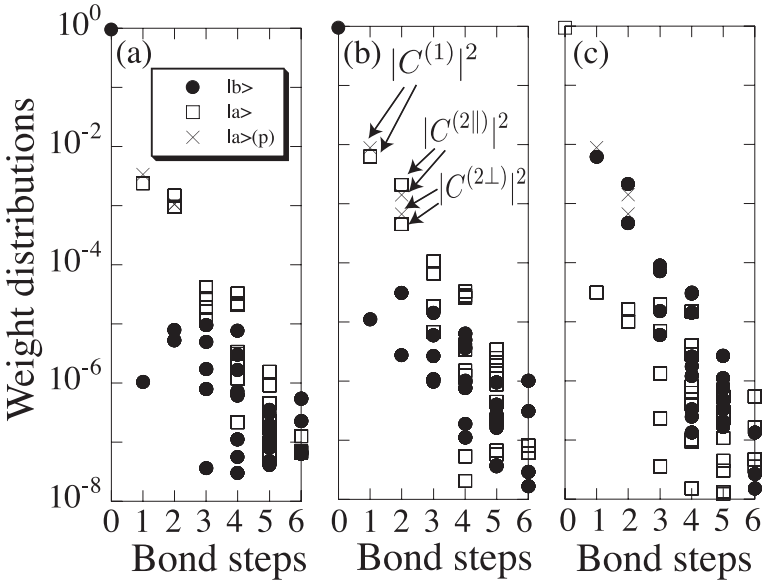


Figure 5.11: Weight distributions of Wannier states on each (anti)bonding orbital, as a function of the bond step from the central bond [27]. The closed circles and the open squares denote the weights on bonding and anti-bonding orbitals, respectively. (a) the Wannier state in the carbon ($d = 0.8d_0$), (b) the Wannier state in the silicon case ($d = d_0$), (c) the Wannier state in the conduction band in the silicon case. The crosses denote the values from the perturbation theory. Note that, in (a), the two values $|C^{(2\|)}|^2$ and $|C^{(2\perp)}|^2$ from the perturbation theory are almost identical.

manner as in Fig. 5.11. The difference from Fig. 5.11 is the fact that Fig. 5.12 shows the exact Wannier state with a smaller periodic cubic cell that contains $N_{\rm A} = 512$ atoms. The above difference of the simulation cell will give no significant difference in the following analysis of the resultant Wannier states. Note that the present exact Wannier state is constructed by the projection method explained above.

Figure 5.12(a)-(c) shows the Wannier states with different values of d; The case (a) is the carbon case ($d = 0.8d_0$), as in Fig. 5.11(a), and the case (b) is the silicon case ($d = d_0$) as in Fig. 5.11(b). The case (c) is a case with a high metallicity ($d = 1.07d_0$), where the band gap is almost vanished (0.15 eV). The case (d) is the silicon case ($d = d_0$), but only two large hoppings (β_0 and β_1) are considered and the other hoppings ($\beta_2, \beta_3, \beta_4$) are ignored. This simpler Hamiltonian is called 'Weaire-Thorpe model' [87, 88]. The case (e) is the Wannier state for the conduction band with the silicon case ($d = 1.0$), as in Fig. 5.11(c). As a total decay profile, Fig. 5.12 also shows the weight of the 'tail'

$$\mathcal{T}(n) \equiv \sum_{\alpha}^{\nu(\chi)\geq n} |\langle\chi|\phi_j\rangle|^2, \tag{5.53}$$

where the summation is done over the (anti)bonding orbitals *outside* the cutoff bond step n. Note that $\mathcal{T}(0) = 1$ and $\mathcal{T}(\infty) = 0$, from its definition. As a measure of the total locality, the typical value of $(1 - \mathcal{T}(3))$ is about 99.5 %, which corresponds to the weight of the local region shown in Fig. 5.7.

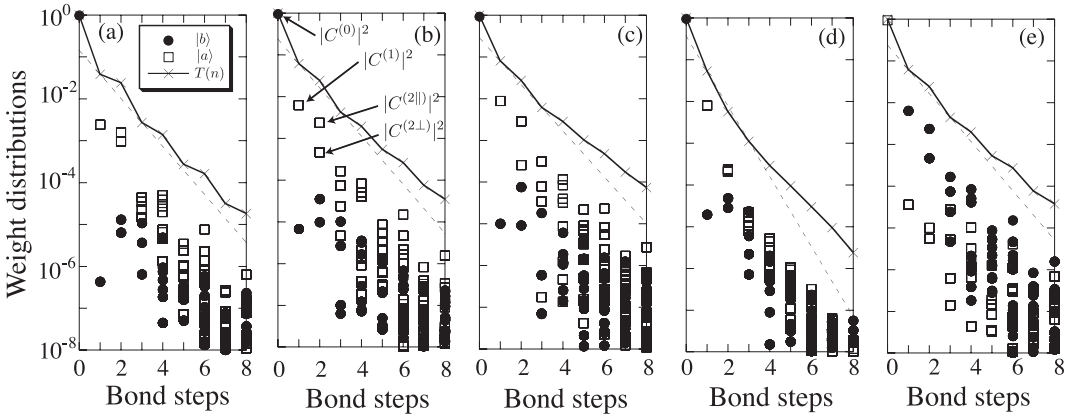


Figure 5.12: Weight distributions of Wannier states on each (anti)bonding orbital and the weight parameter $\mathcal{T}(n)$, as a function of the bond step from the central bond. The closed circles and the open squares denote the weights on bonding and anti-bonding orbitals, respectively. The solid line is $\mathcal{T}(n)$ and the dashed line is the line connecting the two points $\mathcal{T}(1)$ and $\mathcal{T}(3)$. The cases (a)-(c) are the Wannier states of the valence band with (a)$d = 0.80d_0, \alpha_{\rm m} = 0.47$, (b)$d = d_0, \alpha_{\rm m} = 0.78$, (c)$d = 1.07d_0, \alpha_{\rm m} = 0.93$. The case (d) is the one with $d = d_0$, *using only* β_0 *and* β_1. The case (e) is the Wannier states in the *conduction* band with $d = d_0$.

Apart from the decay property, Fig. 5.11 and Fig. 5.12 show the following two properties of the Wannier states, except Fig. 5.12(d); (i) The values of the coefficients of the second nearest neighbor bond sites, $C^{(2\|)}$ and $C^{(2\perp)}$, are split. (ii) The weight distributions show zigzag lines as the function of the bond steps. This is because the spatial extension of the Wannier states is contributed by *two* hopping mechanisms, that is, the nearest neighbor hopping and the second nearest neighbor hopping. The two hopping mechanisms are seen directly in the perturbative formulation, as the first and second terms of Eq. (5.26). Here we recall that the ratio between the interatomic hopping integrals $(\beta_1, \beta_2, \beta_3, \beta_4)$ are almost unchanged among the elements, due to the universality. Within the above tendency, the two hopping mechanisms are quite different in the dependence of the metallicity $\alpha_{\rm m}$. The nearest neighbor hopping is dependent on the metallicity, which appears as $(\alpha_{\rm m}/8)$ in Eqs. (5.25),(5.26), because it is determined by the ratio between the interatomic hopping β_1 and the intraatomic hopping β_0 (See Eq. (5.25)). The second nearest neighbor hopping, on the other hand, is independent on the metallicity $\alpha_{\rm m}$, because it is determined by the ratio between two interatomic hoppings (See Eqs. (5.27),(5.28)). Therefore, the perturbative coefficients of Eqs. (5.25),(5.26) are reduced to

$$\frac{C^{(1)}}{C^{(0)}} \approx \frac{\alpha_{\rm m}}{8} \tag{5.54}$$

$$\frac{C^{(2\|)}}{C^{(0)}} \approx \frac{1}{34} + \left(\frac{\alpha_{\rm m}}{8}\right)^2 \tag{5.55}$$

$$\frac{C^{(2\perp)}}{C^{(0)}} \approx -\frac{1}{28} + \left(\frac{\alpha_{\rm m}}{8}\right)^2 . \tag{5.56}$$

The above two hopping mechanisms explain the properties (i) and (ii); (i) The sum

of the two terms in Eq. (5.55) or Eq. (5.56) directly gives the splitting in the second nearest neighbor coefficients. (ii) The existence of the two hopping mechanisms explains the zigzag line of the weight distribution.

In Fig. 5.12(d), the hopping integrals β_3, β_4 are ignored, which means the lack of the second nearest neighbor hopping. In the perturbative formulation, the first term of Eq. (5.55) or Eq. (5.56) should be replaced by zero. The resultant Wannier state does not show the above two properties (i) (ii). It should be noted that the parametrization in Fig. 5.12(d) does not satisfy the universal tight-binding theory.

Wannier states among different elements (2)

Now we discuss the relation between the band gap Δ and the structure of the Wannier state. In Section 3.1, we explained that the band gap can be estimated as Eq. (3.11). Here the result of the estimation is rewritten;

$$\Delta_{\rm est} \equiv 2|\beta_1|\,(1-\alpha_{\rm m}). \tag{5.57}$$

A negative value of $\Delta_{\rm est}$ means the band overlap between the valence and conduction bands. In Fig.5.13, several results are plotted as the function of the estimated band gap $\Delta_{\rm est}$. The plotted quantities are the band gap Δ, the metallicity $\alpha_{\rm m}$ and the first nearest neighbor coefficient $|C^{(1)}|/|C^{(0)}|$. In the cases with $\Delta = 0$, the wave functions are generated by the above-discussed projection method with Eq. (5.50). These results shows an extrapolation of the cases with nonzero band gaps ($\Delta > 0$).

As an entire tendency in Fig. 5.13, we observe that the band gap Δ decreases with the increase of $\alpha_{\rm m}$ ($\alpha_{\rm m} \to 1$), as is expected from Eq. (5.57). We also observe a good agreement between $\alpha_{\rm m}$ and $8|C^{(1)}/C^{(0)}|$, as is expected from Eq. (5.54). Particularly, the band gap vanishes at $|C^{(1)}/C^{(0)}| \approx 1/11$, which is close to the expected value 1/8.

It is important that, even if the band gap almost vanishes, the dominant weight of the Wannier state is still localized within the central bond, which justifies the perturbative treatment. This can be explained by the two hopping mechanisms; The nearest neighbor hopping contributes the nearest neighbor coefficient $C^{(1)}$ as $\alpha_{\rm m}/8$. The sum of the weight among the six nearest neighbor bond sites ($\mathcal{W}_1$) can be estimated as

$$\mathcal{W}_1 \approx 6\left(\frac{\alpha_{\rm m}}{8}\right)^2. \tag{5.58}$$

Due to the presence of the factor 1/8, the weight $\mathcal{W}_1$ in Eq. (5.58) is still small, even if the band gap becomes small ($\alpha_{\rm m} \leq 1$). The second nearest neighbor hopping is almost unchanged among elements, which is seen as 1/34 or $-1/28$, in Eq. (5.55) or Eq. (5.56), respectively. In results, the coefficients of the second nearest neighbor bond sites are not significantly changed with the change of the band gap, which is seen in Fig. 5.12(a)-(c).

The above explanation shows that the factor 1/8 in Eqs. (5.54), (5.55), (5.56), plays a crucial role of the locality of Wannier states. The origin of the factor 1/8 is given by the factor 1/4 in Eq. (5.17). The origin can be clarified as follows; If two hybridized orbitals are formed in one atom with the sp hybridization

$$|{\rm h}_1'\rangle \equiv \frac{|{\rm s}\rangle + |{\rm p}_z\rangle}{\sqrt{2}}$$

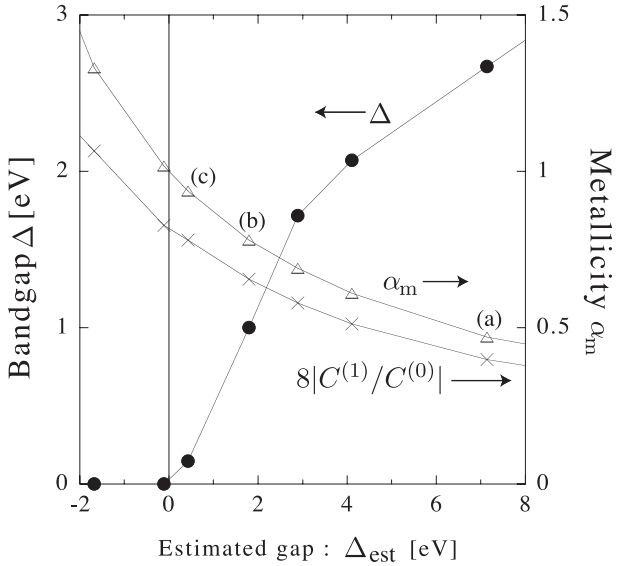


Figure 5.13: The quantities Δ, $8|C^{(1)}/C^{(0)}|$ and $\alpha_{\rm m}$ are plotted as a the function of the estimated band gap $\Delta_{\rm est}$ [85]. The exact calculation is carried out using the periodic cell with 512 atoms. The characters (a)-(c) indicate the the cases in Figs. 5.12 (a)-(c), respectively.

$$|{\rm h}_2'\rangle \equiv \frac{|{\rm s}\rangle - |{\rm p}_z\rangle}{\sqrt{2}}, \tag{5.59}$$

the corresponding intraatomic hopping is reduced to

$$\langle {\rm h}_1'|H|{\rm h}_2'\rangle = \frac{1}{2}\left\{\langle {\rm s}| + \langle {\rm p}_z|\right\} H \left\{|{\rm s}\rangle - |{\rm p}_z\rangle\right\} = -\frac{\varepsilon_{\rm p} - \varepsilon_{\rm s}}{2}. \tag{5.60}$$

When we compare the above result and Eq. (5.17), we can find that the factor 1/4 in Eq. (5.17) is directly related to the sp^3 hybridization. In other words, the factor 1/4 is directly related to the four-fold coordination or a three-dimensional effect. The above discussion shows that the locality of the present composite band Wannier states is directly related to the mixing freedom of the bands. We have discussed that the locality of the composite band Wannier states can be explained, generally and quantitatively, as the virtual impurity state (See Section 4.2). We should say that it may *not* be fruitful to try to understand their locality within the analogy to that of the conventional (isolated band) Wannier states.

Finally, we comment on the the long-distance, or 'tail', behavior of the Wannier state. In the present context, the locality is discussed in the sense that the dominant weight is occupied at a few bond sites. The long-distance behavior seems to be sensitive to the value of the band gap. See, for example, the weight distribution at the bond step $n = 8$ in Fig. 5.12 (a)-(c). Such long-distance behavior, however, does not explicitly contribute to the energy, which is the fundamental justification of the order-N methods, as discussed in Section 2.4.

Part III

Application to fracture of nanocrystalline silicon

Chapter 6

Backgrounds

6.1 Fracture theory and silicon

For the present understanding of fracture mechanism, a pioneering work was given, in the 1920's, by Griffith [89] within a continuum theory. Nowadays, many industrial manufacturing processes involve some fracture processes and many material designs are based on the techniques against fracture. There are a huge number of the related investigations on various materials with various purposes. See Refs. [90, 91, 92, 93] for an overview. Appendix C.2 is prepared as a brief review of the fundamental continuum theory of brittle fracture.

Silicon is an ideally brittle material and is studied intensively, because we can obtain essentially dislocation-free single crystals. Several fundamental mechanisms, such as the brittle-ductile transition [94, 95, 96], are investigated in silicon as an ideal material. In macroscale samples, the easiest cleavage plane is the (111) plane, in which the surface structure forms the 2×1 structure [97, 98, 99, 100]. The 2×1 structure is stable at room temperature but will be transformed irreversibly into the famous 7×7 structure at high temperatures. In other words, the 2×1 structure is a metastable structure.

Here we summarize the Griffith theory [89] (See Appendix C.2 for details). As an ideal situation, suppose a two-dimensional case shown in Fig. 6.1, in which the macroscale sample has an initial crack with the length of $2c$ and is under the uniaxial external load σ. The total energy is described by the energy competition between the energy gain of the strain relaxation and the loss of the surface formation energy. The former energy is a volume term that is proportional to $(\text{length})^3$, while the latter energy is a surface term that is proportional to $(\text{length})^2$. The dimensional analysis gives a typical length scale, which is analogous to the theory of nucleation, (See textbooks of statistical mechanics, such as Ref.[101]). In the present case, the critical length is given as

$$c = c_{\rm G} \equiv \frac{2}{\pi}\frac{\gamma E}{\sigma^2}. \tag{6.1}$$

The derivation is given in Appendix C.2. The quantity γ is the loss of the surface formation energy per unit surface area. The quantity E is the Young modulus within the two-dimensional cases; $E = E_{\rm 3D}$ in plane stress ('thin' plates) or $E = E_{\rm 3D}/(1-\nu^2)$ in plane strain ('thick' plates), with the ordinary Young modulus $E_{\rm 3D}$ and the Poisson ratio ν. Equation (6.1) gives the critical crack length $c_{\rm G}$ for the spontaneous fracture propagation. In other word, if the sample contains a crack with the length of c, the critical stress for fracture σ will be determined by Eq. (6.1).

The above Griffith theory gives a consistent and quantitative picture for the fracture mode of the Si(111) plane with the surface formation energy of $\gamma \approx 1[\text{J/m}^2]$ [102]. Here the surface formation energy was estimated from several experimental results and electronic structure calculations [102] (See Appendix C.2). The (110) plane, another possible cleavage plane in macroscale samples, was also investigated [103, 104], in which the corresponding surface energy γ is obtained on the same order. To convert physical quantities into an atomistic scale, a typical atomistic length is introduced

$$d_0 \equiv \sqrt[3]{v_0} \approx 3[\text{Å}], \tag{6.2}$$

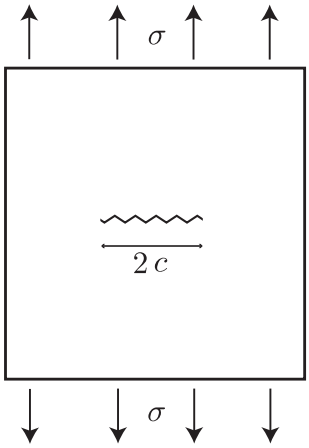


Figure 6.1: Schematic picture of a sample with a crack under the external load σ. The length of the crack is defined as $2c$.

where v_0 is the volume per atom in bulk silicon. The length d_0 is comparable to the bond length (2.35 Å). The above surface formation energy ($\gamma \approx 1[\mathrm{J/m^2}]$) is interpreted as a 'chemical' energy per atom, which is on the order of

$$\varepsilon_{\mathrm{chem}} \equiv \gamma d_0^2 \approx 1 \ [\mathrm{eV}]. \tag{6.3}$$

The above energy scale can be reduced to the bond breaking energy. A comparable energy quantity is the heat of fusion (50 kJ/mol $\approx$ 0.5 eV/atom). On the other hand, an external load of

$$\sigma \approx 1 \ [\mathrm{GPa}] \tag{6.4}$$

can be transformed into the averaged strain energy per atom, which is on the order of

$$\varepsilon_{\mathrm{strain}} \equiv \sigma d_0^3 \approx 10^{-1} \ [\mathrm{eV}]. \tag{6.5}$$

For fracture phenomena, the above two energy quantities should satisfy the inequality

$$\varepsilon_{\mathrm{strain}} \ll \varepsilon_{\mathrm{chem}}. \tag{6.6}$$

If the above inequality were in the reverse order ($\varepsilon_{\mathrm{strain}} \gg \varepsilon_{\mathrm{chem}}$), the energy increase due to the external load would exceed the *total* binding energy and *all* the chemical bond would be broken. With the above quantities, Eq. (6.1) is rewritten as the product of dimensionless quantities;

$$\begin{aligned} \frac{c_{\mathrm{G}}}{d_0} &= \frac{2}{\pi} \times \frac{E}{\sigma} \times \frac{\gamma d_0^2}{\sigma d_0^3} \\ &= \frac{2}{\pi} \times \frac{E}{\sigma} \times \frac{\varepsilon_{\mathrm{chem}}}{\varepsilon_{\mathrm{strain}}} \end{aligned} \tag{6.7}$$

For a silicon case, the critical length $c_{\rm G}$ is estimated. The Young modulus E is on the order of 10^2 GPa. The critical crack length with $\sigma = 1$[GPa] is given as

$$c_{\rm G} \approx \frac{E}{\sigma} \times \frac{\gamma}{\sigma} \approx \frac{10^2\,[\mathrm{GPa}]}{1\,[\mathrm{GPa}]} \times \frac{1\ [\mathrm{J/m^2}]}{1\,[\mathrm{GPa} = 10^9 \mathrm{J/m^3}]} \approx 100\,[\mathrm{nm}]. \tag{6.8}$$

Since the critical length $c_{\rm G}$ is proportional to σ^{-2} ($c_{\rm G} \propto \sigma^{-2}$), the critical length with $\sigma = 10$[MPa] will be a macroscale length

$$c_{\rm G} \approx 1[\mathrm{mm}]. \tag{6.9}$$

To see a typical scale of the macroscale fracture experiment, we pick out a recent fracture experiment [105], in which the sample size is $150\,\mathrm{mm} \times 100\,\mathrm{mm} \times 0.75\,\mathrm{mm}$ and the critical stress is given by $\sigma = 5 - 15$ MPa.

In short, the above case of the Si(111) cleavage plane is an case that the theoretical and experimental results give a consistent picture within the Griffith theory. One unsettled issue is how and why the cleaved surface is formed in the (111) surface with the metastable (2×1) reconstructed structure. As well as the above issue, we focus on the fracture of *nanocrystals*, which gives the purpose of the simulations in Chapter 7.

6.2 Requirement on atomistic theory for fracture

In this section, the requirement for atomistic theory is discussed in the context of the fracture simulation of silicon (nano)crystals. Especially, we will compare classical models and electronic structure calculations. As discussed in the previous section (Section 6.1), the total (static) energy for fracture phenomena should be written as

$$\begin{aligned} & \text{(total energy)} \\ = \; & \text{(energy for strain relaxation in bulk region)} \\ + \; & \text{(energy for structural changes in surface and/or crack tip region)}, \end{aligned} \tag{6.10}$$

where we describe the energy contributions from the crack tip region and the other surface region as one energy term (the second term). In the continuum theory, the crack tip region is distinguished from the other surface region, due to the singularity of the stress fields (See Appendix C.2). In the atomistic theory, on the other hand, there is no reason to distinguish the crack tip region from the other surface region. In Eq. (6.10), the first energy term originates from small deformation in the bulk region, which can be described by the theory of linear elasticity. The second energy term originates from bond breaking and rebonding, which is beyond linear elasticity. Therefore, the essential requirement for the atomistic theory is to reproduces the two energy terms in Eq. (6.10).

There are several popular classical potentials for silicon, such as the Stillinger-Weber potential [82] and the Tersoff potentials [106, 107]. See Ref.[108] for the comparison between classical potentials. Using classical potentials, several molecular dynamics simulations have been performed for fracture simulation of silicon crystals. As a recent one, we pick out a study with 10^5 atoms [109, 110]. A more recent work, however, pointed out the limited applicability of classical modelings and the importance of the electronic structure calculations [103, 104]. As a fundamental point, it is difficult to construct an unique classical model that can reproduce the atomic structures for various 'environments' of bulk and non-bulk phases. One of the major difference for the 'environment' is the difference of the coordination number of atoms, which is the crucial difference between the first and second energy terms in Eq. (6.10). The above difficulty was pointed out, for instance, in the original paper of the Tersoff potential [107], in which two parameter sets, called 'Si(B)' and 'Si(C)', were prepared for different environments. In the above context, a dynamical fracture simulation is one of the severest application, because the environment of atoms is dynamically changed from the bulk one to non-bulk ones.

As explained in Chapter 3, the difference of 'environments' is directly related to the quantum mechanical freedoms of electronic structures. In Chapter 3, we have explained how the quantum mechanical freedoms govern the atomistic structures among solid, liquid and surface phases. These structures are systematically understood by the universal tight-binding theory. For example, the asymmetric dimer of the Si(001) surface is governed by the following quantum mechanical freedoms; (i) the hybridization freedom between s and p orbitals, (ii) the orthogonality relation between wave functions. In the present order-N method, the above two freedoms are included, which is the main reason why our order-N calculation reproduces the

asymmetric dimer of the Si(001) surface. We will see, in the next chapter (Chapter 7), these quantum mechanical freedoms are essential in the fracture processes.

Now we discuss possible theoretical connections between the electronic structure calculation and the classical models. In Section 4.3, we discussed the perturbative formulation with non sp^3 wave functions as a key for classical models in non-bulk phases. Such perturbative formulations will be possible, if reliable unperturbed wave functions are prepared. For example, the different models will be prepared in bulk and surface regions, by preparing different unperturbed wave functions. The applicability of the models is justified, when the calculated perturbative terms are much smaller than the unperturbed term, which can be checked during the simulation. The application of the above theoretical approach may be useful for fracture simulations, but is beyond the present thesis.

Finally, we should say that the atomistic simulation with macroscale number of atoms (10^{23} atoms) is impractical, even with the present large-scale electronic structure calculations and with parallel computers. Therefore, the atomistic theory of fracture should be reasonably connected to the continuum theory, when we would like to discuss the fracture phenomena from nanoscale to macroscale. We will discuss this point later, in the summary of our simulation results (Section 7.7).

6.3 Energetics of Si(001) surface

The Si(001) surface is important in the general context of the nanoscale material theory, since it is used as the standard template in the present semiconductor technology. Here we review the energetics of the related structures, because the Si(001) surface will appear in the fracture simulation in Chapter 7. As already explained in Section 3.2, the asymmetric dimers are formed as the basic structure in the Si(001) surface. In this section, we will discuss (i) flipping freedoms of the asymmetric dimers and (ii) step structures.

Flipping freedoms of the asymmetric dimers

In experimentally observed Si(001) surfaces, the dimer alignment forms rows, as in Fig. 6.2. Here the figure (a) shows the dimer rows of the asymmetric dimer in the alternately buckled configuration, which is denoted as the '(4×2)' configuration. The figure (b) shows, on the other hand, the dimer row of the *symmetric* dimers, which is denoted as the '(2×1)sym' configuration. Now the 'intra-row' direction is denoted as the $[1\bar{1}0]$ direction in Fig. 6.2, while the 'inter-row' direction as the [110] direction.

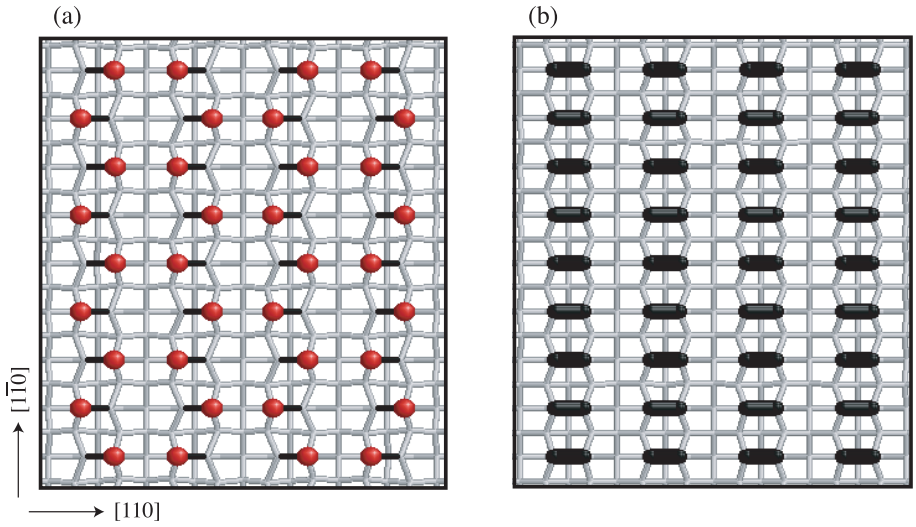


Figure 6.2: Geometry of the Si(001) surface with asymmetric dimers (a) and symmetric dimers (b). In (a), a black rod shows the reconstructed bond and a red ball shows the upper atom of the asymmetric dimer. In (b), a bold black rod shows the reconstructed *double* bond. The former one is in the (4×2) configuration with respect to the flipping freedom. The geometries are obtained in the calculation of Fig. 3.2.

Now the fundamentals of the electronic structure are explained for the Si(001) surface. One surface dimer has, formally, four dangling bond orbitals, as explained in Section 3.2. Among them, two orbitals are transformed into the σ bonding state and the σ^* antibonding states, which lie among the two atoms of a surface dimer. The energy levels of the bonding and antibonding states lie within the energy region

of the valence and conduction bands, respectively. The other two dangling bond orbitals are transformed into the two surface states, which are denoted as the 'π' and 'π^*' states. The 'π' state, the lower energy state, is occupied and its physical picture has been explained in Section 3.2. If the 'π' state is a π-bonding state in a symmetric dimer, the 'π^*' state is the corresponding antibonding state. If the 'π' state is an atomic state of the 'up' atom in an asymmetric dimer, the 'π^*' state is another atomic state of the 'down' atom. The corresponding energy levels forms the surface band that appear within the energy gap between the valence and conduction bands. As a common feature in Fig. 6.2(a) and (b), the numbers of bond steps between two nearest neighbor surface dimers are different among the intra- and inter-row directions. In the *intra-row* direction, two nearest neighbor dimers lies with *three* bond steps, while, in the *inter-row* direction, two nearest neighbor dimers lies with *five* bond steps. From the viewpoint of the nearest neighbor tight-binding Hamiltonian, two nearest neighbor dimers are coupled with the successive hopping along *three* bond steps in the *intra-row* direction, but along *five* bond steps in the *intra-row* direction. This explanation should result in an anisotropic dispersion of the surface band along the intra- and inter-row directions. The electronic structure calculation with a tight-binding Hamiltonian can be seen, for example, as Fig.1 of Ref. [73], which shows the expected anisotropic dispersion of the surface band; The surface band shows a large energy dispersion in the *intra-row* direction, such as $J \to K$, but shows a small dispersion in the *inter-row* direction, such as $K \to J'$. This tendency can be also found in *ab initio* calculations, such as Fig.19 in Ref. [66]. This fact is consistent to the above explanation with the number of bond steps. Here two technical comments are added; (i) The electronic structure figures in Ref. [73] and Ref. [66] are drawn in different definitions of the unit cell. See these papers for details. (ii) *Ab initio* electronic structure calculations may be carefully discussed for surface bands, because the LDA usually underestimates the band gap. See the last paragraph of Appendix A.1.

With respect to the flipping freedom of the asymmetric dimers, there are several possible configurations, which are schematically shown in Fig. 6.3. Here an arrow indicate the buckling of the dimer. In other words, an arrow is a projected vector from the 'down' atom into 'up' atom in a dimer. The figure(a) shows the '(4×2)' configuration, the same structure as that in Fig. 6.2(a). The figure (b) and (c) are denoted as the '(2×2)' and '(2×1)' configuration, respectively. Sometimes the flipping of the asymmetric dimers is formally mapped to the flipping of virtual 'spins'. In terms of the spin structure, the '(4×2)' configuration is the anti-ferro configuration both in the inter- and intra-row directions. The '(2×2)' is the anti-ferro configuration in the intra-row direction but the ferro configuration in the inter-row direction. The '(2×1)' configuration is the ferro configuration both in the inter- and intra-row direction. The results of electronic structure calculations are summarized in Table 6.1, which shows the '(4×2)' configuration is the ground state of the perfect (001) surface. The energy differences in the table can be understood by the difference of the coupling constants between the flipping freedoms among dimers. From the table, we observe the following properties; (i) the energy difference between the (4×2) and (2×2) configurations is smaller than that between the (2×2) and (2×1) configurations. This fact is consistent to the above explained fact that the *inter-row* coupling is described by the hopping in larger number of bond steps (five bond steps)

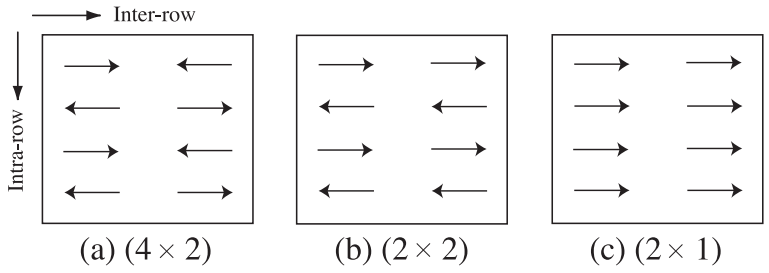


Figure 6.3: Schematic pictures of the reconstructed Si(001) surfaces. The flipping freedom of an asymmetric dimer are indicated as an arrow from the lower atom to the upper atom; (a) (4×2) configuration, (b) (2×2) configuration and (c) (2×1) configuration.

(meV/Dimer)	order-N	exact TB	*ab initio*
$E_{(2\times1)} - E_{(2\times2)}$	94.2	62.4	48 ± 18
$E_{(2\times2)} - E_{(4\times2)}$	18.0	1.2	3 ± 13

Table 6.1: Calculated energy difference among the different configuration of the flipping freedom on Si(001) surface. The energies are calculated by the variational order-N method (present work), an exact tight-binding (TB) calculation [111] and an *ab initio* calculation [66].

than the *intra-row* coupling (three bond steps). (ii) The order-N method results in a large error from the exact (diagonalization) results. The methodological problem, however, is not only within the order-N method, since Table 6.1 contains the energy difference in a quite fine scale (meV/atom). Such an energetically delicate structure should be carefully discussed, generally, in the total energy methods, which is implied by the large error bars of the *ab initio* result in Table 6.1.

Here we discuss the dynamics of the flipping freedoms. At room temperature, the flipping freedoms is not frozen and an STM image is obtained with *symmetric* dimers, due to the time average of the flipping motions. Low temperature STM observations give images with the asymmetric dimers [67]. The above experimental fact is consistent to the results in Table 6.1, which contains the energy scale smaller than the kinetic energy of room temperature ($300\,\mathrm{K} \approx 1/40\,\mathrm{eV}$). A direct theoretical approach was done with a model 'spin' Hamiltonian [112, 113], which maps the flipping motion, formally, to that of a two-dimensional 'spin' system. The 'spin' Hamiltonian contains the spin-spin interaction parameters and the parameters are determined so as to reproduce the calculated total energies among different 'spin' or flipping configurations. As another experimental feature, the atomic scale local environment, such as surrounding defects or steps, seriously influences the flipping motion of the dimers. See, for example, the introduction part of a related theoretical paper [111]. In the paper, the flipping motion on perfect or defective Si (001) surfaces is studied by tight-binding molecular dynamics with 10^2 atoms. The flipping motion was observed experimentally [114], by the time trace of the STM tunneling current at an atom of flipping dimer with $T = 70$ K. The sampling rate is 16.7 kHz ($\approx (0.06\,\mathrm{ms})^{-1}$) and the observed frequency of the flipping motion is 0.81 kHz

($\approx (1.2\,\mathrm{ms})^{-1}$). In short, the flipping freedom of the asymmetric Si(001) surface dimers can be thermally driven and is sensitive to local atomic scale environments. Its dynamics is important, at least, for its influence on various processes, such as epitaxial growth, etching, and chemical reaction.

Step structures and surface strain energy

Experimentally observed Si(001) surfaces contain several step structures and now their energetics is focused. See a review [115]. The four types of step structures can be defined in the Si (001) surface. They are denoted as S_A, S_B, D_A and D_B [116], and are shown schematically in Fig. 6.4. Here the letter 'S' or 'D' indicates the single (S) or double (D) step structure. The corresponding atomistic pictures can be seen in Refs.[116, 115] or textbooks.

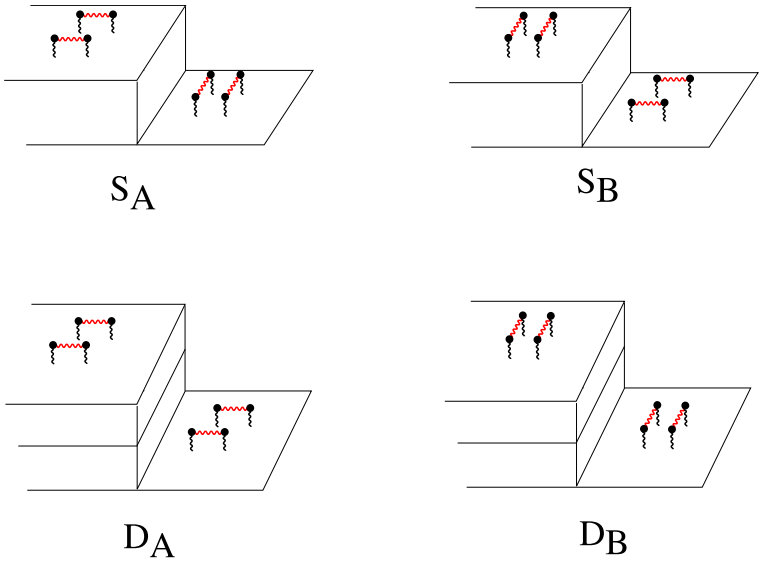


Figure 6.4: Schematic pictures of step structures on the (001) surface. The four types of step structures are denoted S_A, S_B, D_A and D_B, respectively, according to Ref. [116]. A red spring indicates the bond of the surface dimer. The asymmetry of the dimer is ignored in the picture. The actual dimers are asymmetric within the alternating flipping geometry.

Within a perfect Si(001) surface, the asymmetric dimers are formed in the one direction , such as the [110] direction in Fig. 6.2. The resultant structure shows the (2×1) symmetry, if the flipping freedoms are ignored. Across the single step structures S_A or S_B, the dimer directions are perpendicular between the upper and lower atomic layers. In other words, the single step structures is a boundary between a (2×1) domain and a (1×2) domain. In the S_A structure, the step lies in the parallel direction to the dimers of the *lower* atomic layer, while, in the S_B structure, the step lies in the parallel direction to the dimers of the *upper* atomic layer. Across a double step structure, D_A or D_B, the dimer directions are parallel between the upper and lower atomic layers. In other words, the double step structures is a boundary between a (2×1) domain and another (2×1) domain on a different atom layer. In

the D_A structure, the step lies in the perpendicular direction to the dimers, while, in the D_B structure, the step lies in the parallel direction to the dimers.

The step formation energies for the four types of steps are summarized in Table 6.2. Among electronic structure calculations in the table, the step formation energy is defined as the loss of the total energy to form the step in the clean surface. An important consequence from the electronic structure calculations is that, within the single steps, the S_A step can be formed easier, or at a smaller energy cost, than the S_B step

$$\lambda(S_A) < \lambda(S_B). \tag{6.11}$$

This result predicts that the S_B step should have many thermally excited kinks that consists of segments of the S_A steps, due to the small energy cost of the S_A step. The S_A step, on the other hand, should not have such kink structures. The above tendency is seen in STM images. As a reverse problem, the step formation energies can be estimated, quantitatively, from the STM images [117]. The results are also shown in Table 6.2. Similar analysis of the STM images is reviewed in Ref.[115].

(eV/2d)	$\lambda(S_A)$	$\lambda(S_B)$	$\lambda(D_A)$	$\lambda(D_B)$
Tight-binding [116]	0.02	0.30	1.08	0.10
Ab initio [118]	0.18	0.24	0.86	0.34
Experiment [117]	0.056	0.18	–	–

Table 6.2: Step formation energies of the Si(001) surface. The upper and middle columns give theoretical values from the electronic structure calculations. The lower column gives the estimated values from the analysis of experimental STM images. $d \equiv 3.84$Å.

Now we discuss that the anisotropic surface strain energy is also important in determination of step structures [119]. The formation of surface dimers causes the anisotropic surface strain in the dimer direction and the perpendicular direction. If double steps appear dominantly, the resultant (001) surface are covered only with the (2×1) domains, which accumulates the anisotropic strain energy. The situation can be schematically shown as

$$\cdot\cdot \quad || \quad (2\times 1) \quad || \quad (2\times 1) \quad || \quad (2\times 1) \quad || \quad \cdot\cdot, \tag{6.12}$$

where the symbol '|| ' denotes a double step. The presence of a *single* step flips the anisotropic direction between the lower and upper layers. The resultant surface is covered with (2×1) and (1×2) domains in the alternately phases between the *single* steps, which will reduce the anisotropic strain energy. The situation can be schematically shown as

$$\cdot\cdot \quad | \quad (2\times 1) \quad | \quad (1\times 2) \quad | \quad (2\times 1) \quad | \quad \cdot\cdot, \tag{6.13}$$

where the symbol '| ' denotes a single step. See Refs.[120, 121] for the quantitative discussion with *ab initio* calculations. The above discussion predicts that two single steps (S_A+S_B) should be energetically favorable than a double step (D_A or D_B), from the viewpoint of the relaxing the anisotropic surface strain energy.

The energy data in Table 6.2 are calculated without the above strain relaxation mechanism. Table 6.2 gives an inequality

$$\lambda(\mathrm{D_B}) < \lambda(\mathrm{S_A}) + \lambda(\mathrm{S_B}) < \lambda(\mathrm{D_A}), \tag{6.14}$$

which predicts that the $\mathrm{D_B}$ step can be formed easier than the two single steps ($\mathrm{S_A}$+$\mathrm{S_B}$). The present prediction from Eq. (6.14) contradicts the previous prediction from the strain relaxation mechanism. In results, the step formations are determined by the competitive mechanism between the relaxation of surface strain energy and the step formation energy. If the step interval, denoted as l, is infinitely long ($l \to \infty$), the energetics will be governed by the surface strain and the two single step should appear. In the opposite limit ($l \to 0$), the energetics is governed by the step formation energy and the double step should appear. Here we can expect a crossover between the step interval l. Experimentally, the step interval can correspond to the misoriented angle θ in the [110] direction from the ideal (001) surface. A theoretical estimation gave the prediction of the crossover at $\theta \approx 1.2^\circ - 2.5^\circ$ [119, 122]. In experiments, with the miscut angle larger than about 1.5°, double steps begin to form and their fraction increases with increasing miscut angle toward a maximum of nearly 100 % at 5-6° [115]. The corresponding STM images of the single-stepped and/or double-stepped surfaces can be seen in many papers, such as Refs. [117, 122, 115].

Here we would like to point out the analogy between the present theory of step structures and the Griffith theory of fracture explained in Section 6.1. The crossover in this section can be expected from the dimensional analysis, because the step formation energy is scaled as (length)^1 and the surface strain energy is scaled as (length)^2. The corresponding critical length is that of the step interval. In the Griffith theory, on the other hand, the critical crack length c_G is obtained by the energy competition between the surface formation energy, scaled as (length)^2, and the bulk strain energy, scaled as (length)^3. The analogy is summarized in Table 6.3. Within the atomistic picture, step formation and surface formation processes can be commonly classified into 'chemical' processes in the sense that they are formed with bond breaking and rebonding processes. The characteristic energy scale is written as $\varepsilon_\mathrm{chem}$, as in Section 6.1, of which values should be in the same order. On the other hand, the energy scale of the strain energy $\varepsilon_\mathrm{strain}$ should be much smaller than the chemical energy ($\varepsilon_\mathrm{strain} \ll \varepsilon_\mathrm{chem}$), though its actual values may be different among the two cases. The dimensional analysis gives a critical length scale as

$$n \equiv \frac{\varepsilon_\mathrm{chem}}{\varepsilon_\mathrm{strain}} \tag{6.15}$$

in the unit of the number of atomic layers. As a common feature, the chemical energy is contributed by a *small* number of atoms with a *large* energy scale, while the strain energy is contributed by a *large* number of atoms with a *small* energy scale. In results, they are competitive in the total energy. For describing the above competition, the atomistic simulation should be done with a sufficiently large system size to contain the critical length of Eq. (6.15). We will discuss this point again in the final chapter of the present thesis (Chapter 8).

	fracture	step formation
Chemical energy ($\varepsilon_{\text{chem}}$)	surface (2D)	step (1D)
Strain energy ($\varepsilon_{\text{strain}}$)	bulk (3D,isotropic)	surface (2D, anisotropic)
Critical length ($n \equiv \varepsilon_{\text{chem}}/\varepsilon_{\text{strain}}$)	crack length	step interval

Table 6.3: The dimensional analysis of the energetics in fracture and step formation.

Chapter 7

Fracture of nanocrystalline silicon

7.1 Purpose

In this section, we describe the purpose of the present fracture simulations of silicon. Here the same notations are used as in Section 6.1. We will focus on the dynamical brittle fracture of nanocrystalline silicon under the external load in the [001] direction. The (001) surface is focused, because of its general importance in the nano technology, as explained in Section 6.3. The results will be analyzed with quantum mechanical freedoms. Especially, we will focus on the following issues;

(I) Dynamical fracture process; how and why the fracture path is formed and propagate in the crystalline geometry. This issue should include the surface reconstruction process.

(II) Fracture behavior of nanoscale samples; its possible difference from the macroscale samples.

Hereafter we discuss the above two issues more clearly. First, the issue (I) is discussed. Since fracture is a thermal non-equilibrium process, the atomic structure on a cleavage surface can be different from that on equilibrium clean surfaces, which was discussed in Section 6.1. Moreover, when we try to construct a *dynamical* description, two limiting pictures can be considered, as shown schematically in Fig. 7.1. In the figure, the surface reconstruction is schematically drawn as surface dimer formations. One limiting picture is the process of (A)→(B)→(C) and the other is the process of (A)→(D)→(E)→(C). In the former process, the bond breaking processes and the surface reconstruction processes occur *simultaneously*. In the latter process, the bond breaking processes ((A)→(D)→(E)) occur *first*, and, *then*, the surface reconstruction processes ((E)→(C)) occur. In other words, the bond breaking processes and the surface reconstruction process are separated in the time scale.

Now we turn to discuss the second issue, the issue for the fracture of nanocrystalline silicon. As explained in Section 6.1, the Griffith theory is a kind of a dimensional analysis and gives the critical crack length for fracture ($c_{\rm G}$). In the above picture, the sample size (L) is larger than the above crack length $c_{\rm G}$ ($L > c_{\rm G}$), as in Fig. 6.1. Since the length $c_{\rm G}$ is not dependent on the sample size L, the fracture behavior can be expected to be different from the above picture in case that the sample size L is smaller than the critical length $c_{\rm G}$ ($L < c_{\rm G}$). In this chapter, we will see such a situation in *nanocrystalline* silicon, in which the numbers of atomic layers for these lengths ($\approx c_{\rm G}/d_0, L/d_0$) are not macroscale numbers. For example (see the next chapter), the critical length is given by $c_{\rm G} \approx 100$nm, while the sample sizes are smaller than 20 nm ($L \leq 20$nm). From the analogy with the theory of nucleation, one can map the above nanoscale situation to the nucleation within a confined space. Now one can expect a crossover among the system sizes between the macroscale and nanoscale sample. Such a crossover has been already discussed in Section 6.3 with the similar dimensional analysis of the theory of step structure in the Si(001) surface.

The difference between nanoscale and macroscale samples can be discussed also from a different viewpoint. In the continuum theory of fracture (See Appendix C.2),

the effect of the crack tip is characterized by the singular stress field

$$\sigma_{ij} \propto \frac{1}{r^2} \quad (r \ll c), \tag{7.1}$$

where $r \equiv \sqrt{x^2 + y^2}$. The point $r = 0$ corresponds to the crack tip. The singular area is given by the assumption of $r \ll c$. The assumption of $r \ll c$ is practical, only when the crack length is much larger than the atomistic length scale ($c \gg d_0$), which may not be expected in nanoscale samples.

Finally, we discuss the initial structure and the boundary condition in the fracture simulations. The choice of these conditions is important, because it will severely restrict the resultant dynamics. In most atomistic simulations [109, 110, 103, 104], simulations are done in 'two-dimensional' samples, in the sense that the simulation cell is periodic in one direction. Moreover, samples have initial well-defined cleavage planes. Unlike these works, we do not impose any periodicity and do not prepare any initial crack plane, so as to discuss the above two issues (I) and (II). We will investigate the fracture of clusters in different sizes, among 10^2-10^5 atoms, with an external load in the [001] direction. We will compare the results among different sample sizes. Though the (001) surface is not the easiest cleavage plane of macroscale samples, macroscale fracture behaviors will appear, as a crossover, if the sample is enough large. The present system sizes, up to 10^5 atoms, seem to be not enough large for the observation of macroscale fracture behaviors, but the present result contains a phenomena that is understood as the beginning of the crossover between nanoscale and macroscale samples. The crossover will be discussed in Section 7.6.

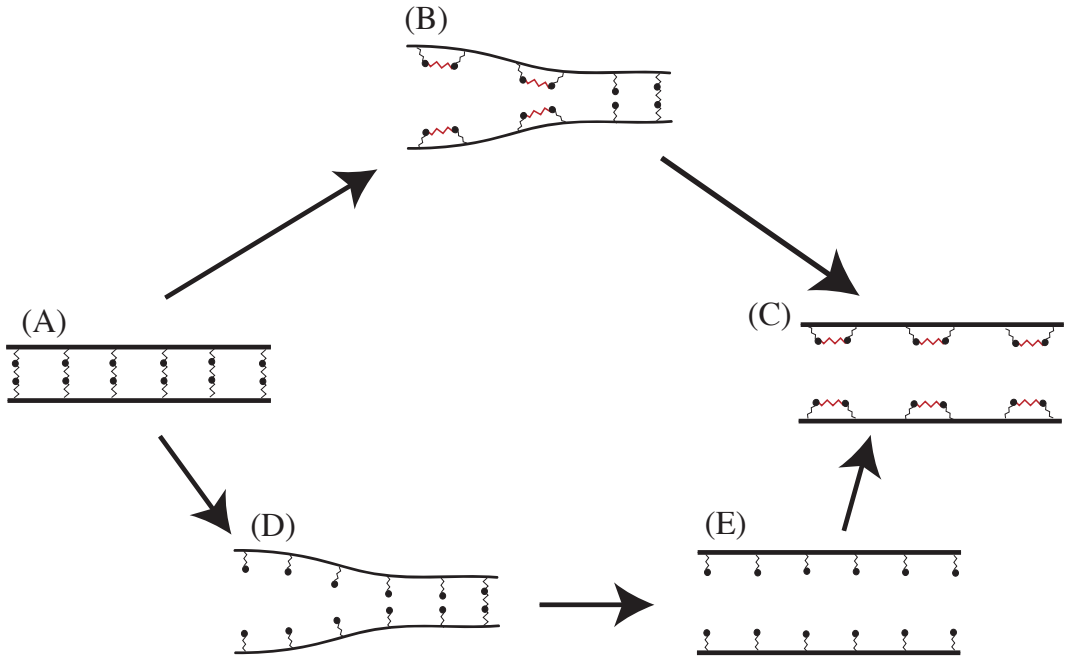


Figure 7.1: Schematic picture of the two possible fracture processes that contains bond breakings and surface reconstructions; (A) initial (crystalline) structure, (B) intermediate structure with reconstructed surface, (C) the final (reconstructed) structure, (D) intermediate structure with unreconstructed surface, (E) the ideal (unreconstructed) structure. The surface reconstruction is schematically shown as dimer formations.

7.2 Basic properties of fracture simulation

Hereafter, in this chapter, we will present fracture simulations of nanocrystalline silicon and will discuss the results. In this section, first, we discuss the conditions of the simulation. Then, the fracture behaviors with a small sample is analyzed. These fracture behaviors are seen almost commonly among *all* the samples in this chapter.

Conditions of simulations

All the samples are isolated tetragonal clusters, whose geometries are labeled with the number of atomic layers in three axes, such as $n_{100}\times n_{010}\times n_{001}$ or $n_{110}\times n_{1\bar{1}0}\times n_{001}$. For fracture propagations, external loads in the [001] direction are imposed. Figure 7.2 shows the schematic picture of a sample. Here we define 'top', 'bottom' and 'side' surfaces of the sample as shown in Fig. 7.2. Note that the sample does not contain an initial crack.

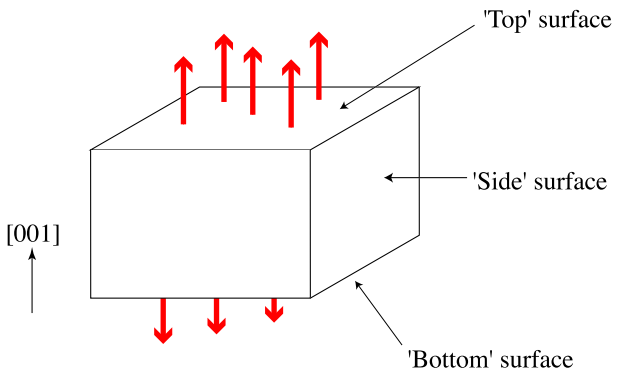


Figure 7.2: Schematic picture of a sample. The external load is shown as red arrows.

The Wannier states at all the sample surfaces are terminated by fixed sp^3 bonding states and are not reconstructed. The time step of the molecular dynamics is 3 fs. During the simulations, the external loads can be dynamically controlled by the atoms on the 'top' and 'bottom' surfaces of the sample. The velocity is chosen, typically, to be 10^{-2} km/s, which is much slower than that of observed fracture propagation velocities (on the order of km/s). The total kinetic energy is controlled to be that with 300 K by the Nosé thermostat method [123, 124]. Note that the thermostat does not affect the essentials of the fracture dynamics, because the thermostat controls the *averaged* kinetic energy among all the atoms in the sample (See Appendix D.4), while the fracture occurs *locally*. Most of the simulations, a defect bond is initially prepared, as a seed of fractures, in a central region of the samples. The origin of the initial defect bond is a short range repulsive potential imposed on one particular pair of atoms

Fracture simulation with 91 atoms

We demonstrate the properties with a small sample, with 91 atoms, and the result is shown in Fig. 7.3. The simulation is done by the exact diagonalization method. Here two simulation details are explained; (i) The bonds in the figure are drawn according to the atomic distance, only for an eye guide. Note that, later in this chapter, we will draw the bonds, due to a quantum mechanical analysis of Wannier states. (ii) Though the sample is quite small, the tetrahedral structure is well preserved, even at the surfaces of the sample This is because the surfaces of the sample are terminated by the Wannier states in sp^3 bonding states. The present boundary condition corresponds to the case in which the system is embedded in a bulk crystal or a tetrahedral sp^3-bonded network. If the surfaces of the sample were terminated by hydrogen atoms, as usually done, the surfaces of the sample would be deformed, due to the deviation from the sp^3-bonded network.

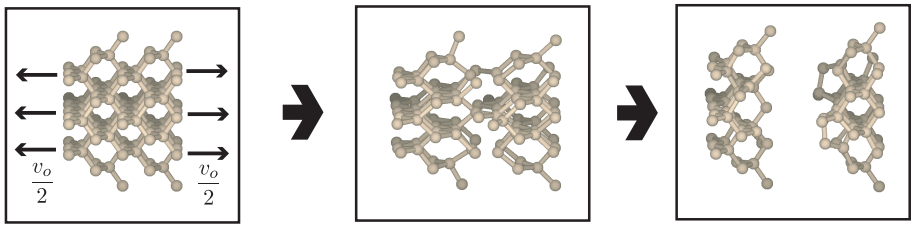


Figure 7.3: Fracture simulation with 91 atoms using the exact diagonalization method. Here the bonds are drawn, just for an eye guide, according to the interatomic distance. The direction of the external load is set as the z axis.

Global property of fracture

Here we discuss the global properties of the fracture. The upper panel of Fig. 7.4 shows the stress at the 'top' or 'bottom' sample surface. The stress is measured as the averaged force among the atoms on the sample surfaces. The lower panel of Fig. 7.4 shows several eigen levels that lie near the highest occupied or lowest unoccupied levels. At the time $t = 0$, we start imposing the external load. First we discuss the stress, the upper panel of Fig. 7.4. The fracture begins at $t \approx 1.5$ ps, when the stress is at its maximum value, and ends at $t \approx 2$ ps, when the stress becomes zero. A small non-zero stress is also seen at $t \leq 0$, because the equilibrium lattice constant in the present small sample is deviated from that in the bulk sample. In the final state, as in the last snapshot of Fig. 7.3, the sample is completely divided into two peaces.

Several important physical quantities appear in the upper panel of Fig. 7.4. (i) The linear region of the stress, $0 \leq t \leq 1.5$ps, should give the Young modulus E_{100}, since the 'top' or 'bottom' surface of the sample is controlled by the constant velocity motion v_0. Though the figure shows a large fluctuation in the stress, we estimated the Young modulus to be $E_{100} \approx 100$GPa, where the estimated value may include an error on the order of 10 %. This value is comparable with the experimental value $E_{100} = 130$ GPa (See B.1), Note that the calculated values of the elastic constants

were discussed in Section 4.3 and the present calculated value should be deviated from the bulk one, due to the small sample size. (ii) The critical stress of $\sigma \approx 2$ GPa gives the critical length $c_{\rm G}$ as $c_{\rm G} \approx 100\,{\rm nm}$, from the discussion in Section 6.1. Within the fracture simulation in this chapter, the fracture begins with the above order of the stress, in which the averaged bond length is about 10 % longer than the equillbrium value. As discussed in Section 6.1, the corresponding strain energy ($\varepsilon_{\rm strain}$) is the order of $\sigma d_0^3 \approx 10^{-1}$ eV, which is smaller than the bond breaking energy ($\varepsilon_{\rm chem} \approx 1$eV). Since the sample length L of the largest sample in this thesis will be $L \approx 20$ nm, the situation in this thesis is that of the nanoscale sample ($L < c_{\rm G}$), as discussed in the previous section (Section 7.1). (iii) The period of the fracture is estimated by $T \approx 2.0 - 1.5 \approx 0.5$ ps. The present sample with 91 atoms gives the sample length of $L \approx 10$ Å. From the above quantities, the crack-propagating velocity $v_{\rm crack}$ is estimated as

$$v_{\rm crack} \approx \frac{L}{T} \approx \frac{10[\text{Å}]}{0.5[\text{ps}]} \approx \frac{10 \times 10^{-10}[\text{m}]}{0.5 \times 10^{-12}[\text{s}]} \approx 2[\text{km/s}], \tag{7.2}$$

which is almost unchanged among the fracture simulations in this chapter. The above value is reasonable, because, as explained in Chapter 6, the experimental value is less than but on the same order of the Rayleigh wave speed ($c_{\rm R} = 4.5$ km/s). The above estimation of the physical quantities shows that the present simulation gives reasonable results.

Here we discuss the electronic structure in the lower panel of Fig. 7.4. In the figure, the highest occupied and lowest unoccupied levels are plotted as the red and blue lines, respectively. They are isolated levels in the band gap and their physical origin is the bonding and antibonding states of the initial defect bond. Hereafter we define 'defect' states, generally, as the electronic states that do not appear in the crystalline structure. The present definition of 'defect' states includes surface states, as well as states in a point defect. During the fracture ($1.5\,{\rm ps} \leq t \leq 2\,{\rm ps}$), the level crossings are seen among several 'defect' levels within the band gap. Especially, the level crossing between the highest occupied and lowest unoccupied levels, red and blue lines in Fig. 7.4, corresponds to the vanishing of the electronic energy gap. This is quite understandable, because the covalent bonding is stabilized by the energy gap between bonding and antibonding states. If the energy gap vanishes, there is no reason to form a bonding state. The above level crossing mechanism should be distinguished from the band overlap between the continuum (valence and conduction) bands. In the final structure, these 'defect' states are transformed into the surface band. Due to the presence of the surface band, the electronic energy gap in the final structure is less than that of the initial crystalline structure. In short, the change of electronic structure in Fig. 7.4(b) is interpreted as the change of 'defect' states, from two isolated states into a surface band. The change in the electronic structure directly corresponds to the fracture from the initial defect bond into an cleaved surface. One may be interested in the simulation *without* the initial defect bond. Such simulation will be discussed in the next section with larger samples.

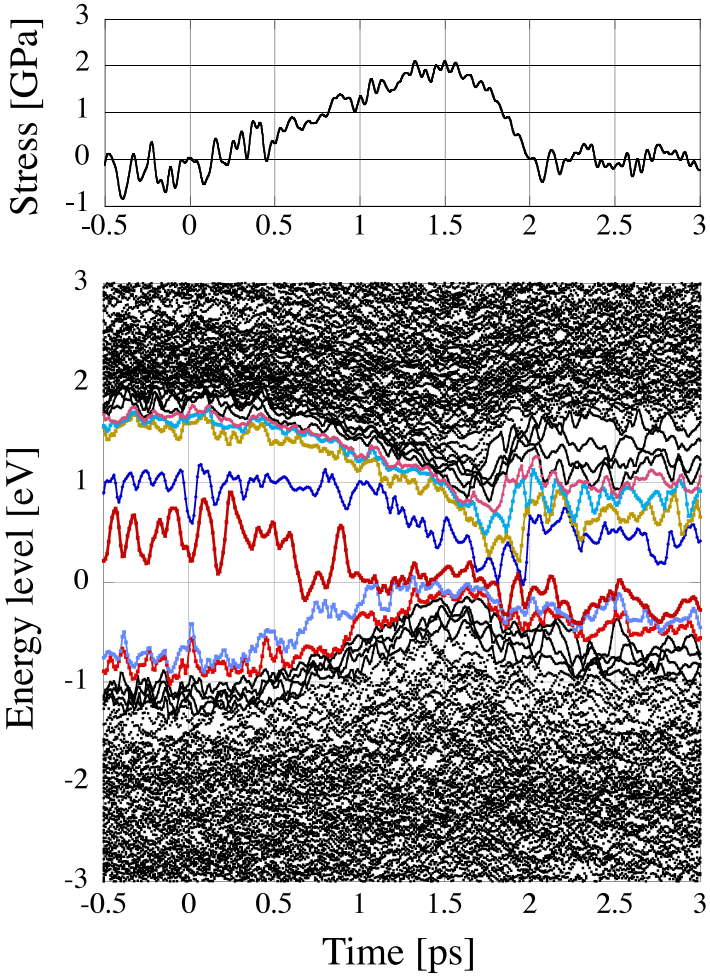


Figure 7.4: Fracture simulation using the exact diagonalization method. The sample contains 91 atoms, of whose geometry is given in Fig. 7.3. The surface stress (upper panel) and several eigen levels (lower panel) are plotted as a function of the time. In the eigen levels, the highest occupied or lowest unoccupied levels are plotted as the red and blue lines, respectively.

Elementary process

Here we describe the elementary fracture process that includes the bond breaking and surface reconstruction processes. Figure 7.5 shows a typical elementary process of a Wannier state $|\phi_i\rangle$, in which we monitor the one-electron energy $\varepsilon_i \equiv \langle\phi_i|H|\phi_i\rangle$ and the weight of s orbitals $f_{\rm s}^{(i)}$ defined in Eq. (3.9). The origin of the time ($t = 0$) is chosen as the time when the energy ε_i has its peak. The time $t = 0$ corresponds to the time of bond breaking, as discussed below. A schematic picture of the process is shown in Fig. 7.6; (a) bulk structure $\rightarrow$ (b) unreconstructed surface $\rightarrow$ (c) reconstructed surface. Small filled circles or solid lines are atoms or bonds that lie in this plane. Small open circles or dashed lines are atoms or bonds that do not lie in this plane. Large open circles are atomic lone pair states. Figure 7.6(d) shows the related energy levels, which is almost the same as in Fig. 5.10. Only one difference between Figs 7.6(d) and 5.10 is the fact that, in Fig. 7.6(d), the sp^3 bonding level indicates the energy of the bulk Wannier state ($\varepsilon_{\rm WS} = -5.08$eV). Before the bond breaking ($t < 0\,$ps), the wave function $|\phi_i\rangle$ is a bonding state in the bulk region, deformed due to the external load. At $t \approx 0\,$ps, a bond breaking occurs; the wave function $|\phi_i\rangle$ loses the bonding character with rapid increase of the bond length. Then ($0\,{\rm ps} < t < 0.2\,{\rm ps}$), the wave function is transformed into a lone pair state localized on one atom, since another bond is broken almost simultaneously at one of the nearest neighbor bond sites. In result, a two-fold coordinated atom appears, as in unreconstructed surfaces, which is shown schematically in Fig. 7.6(a)$\rightarrow$(b). Within the above process ($0\,{\rm ps} < t < 0.2\,{\rm ps}$), the increase of $f_{\rm s}^{(i)}$ ($0.6 \rightarrow 0.8$) causes the energy gain estimated to be $-0.2 \times (\varepsilon_{\rm p} - \varepsilon_{\rm s}) \approx -1.3$eV, which explains the energy gain in the figure ($\varepsilon_i = -2.7{\rm eV} \rightarrow -3.8{\rm eV}$). This can be classified into a dehybridization process, as explained in Section 3.2. Finally, after the thermal motions with a finite time ($t \approx 0.4$ps), a pair of two-fold coordinated atoms forms an asymmetric dimer with a σ bonding state $|\phi_i\rangle$, which is shown schematically in Fig. 7.6(b)$\rightarrow$(c). The resultant asymmetric dimer was discussed in Section 3.2. The corresponding covalent-bonding energy, defined in Eq. (3.20), is $\Delta\varepsilon_i^{\rm (cov)} \approx -1.9$ eV. This energy explains the gain in the figure ($\varepsilon_i = -3.8{\rm eV} \rightarrow -4.8{\rm eV}$) and the energy *loss* (about 1.3eV) due to the decrease of $f_{\rm s}^{(i)}$ ($0.8 \rightarrow 0.6$). This asymmetric dimer is preserved until the end of the simulation, during a couple of pico seconds. In conclusion, the reconstruction process is decomposed into two stages; (i) the formation of two-fold coordinated atoms or an *unreconstructed* surface, (ii) the formation of a *reconstructed* surface dimer. The above two-stage reconstruction process is commonly observed in the present fracture simulations.

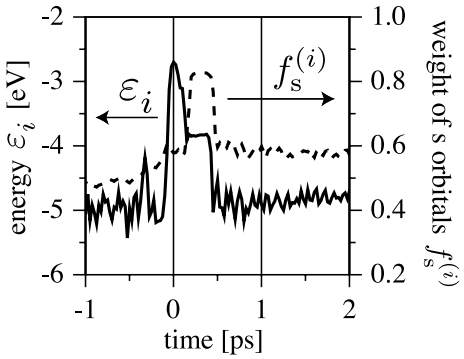


Figure 7.5: Wannier state in the elementary bond breaking and surface reconstruction process; The one-electron energy $\varepsilon_i \equiv \langle\phi_i|H|\phi_i\rangle$ and the weight of s orbitals $f_s^{(i)}$ are plotted as functions of time.

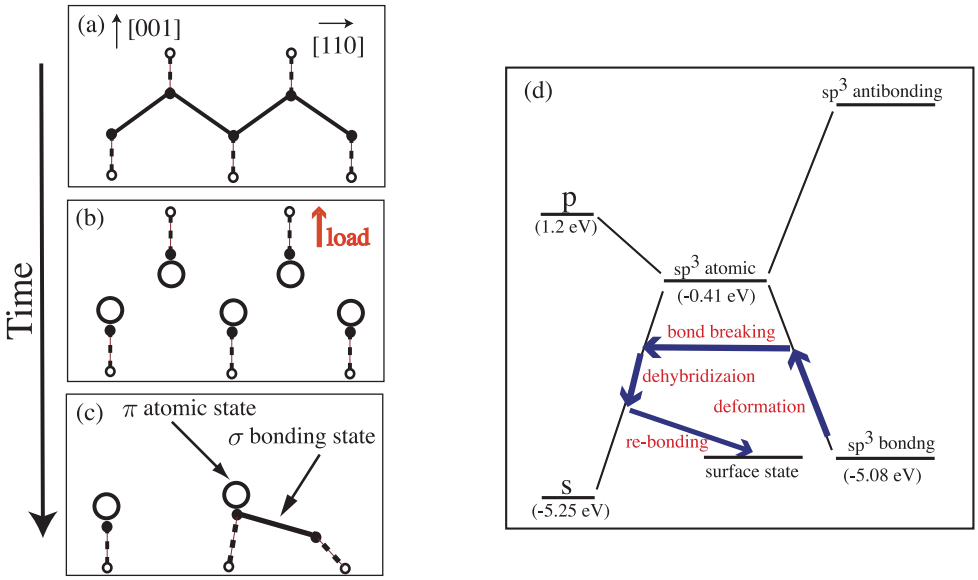


Figure 7.6: (a) (b) (c) The schematic pictures in the elementary bond breaking and surface reconstruction processes; (a) crystal structure, (b) unreconstructed surface and (c) reconstructed surface. (d) Several energy levels of the Wannier state (See Fig. 5.10).

Figure 7.7 shows the calculation of the total density of state (DOS) $D_{\rm tot}(\varepsilon)$ and the partial density of states $D_i(\varepsilon)$ for the above-discussed Wannier state $|\phi_i\rangle$. In Fig. 7.7, the snapshot (a) is that before the bond breaking ($t = -0.1$ps) and (b) is that just on the bond breaking time ($t = 0$). The total and partial DOS are defined as

$$D_{\rm tot}(\varepsilon) \equiv -\frac{1}{\pi}\lim_{\varepsilon_0\to 0}{\rm Im\,Tr}\left[\frac{1}{H+i\varepsilon_0}\right] \tag{7.3}$$

$$D_i(\varepsilon) \equiv -\frac{1}{\pi}\lim_{\varepsilon_0\to 0}{\rm Im}\langle\phi_i|\frac{1}{H+i\varepsilon_0}|\phi_i\rangle. \tag{7.4}$$

The practical calculations are done by the explicit matrix inversion with a finite value of ε_0. In result, the total DOS profile changes its character with the creation of a 'pseudo gap' in the energy region slightly higher than the chemical potential. This corresponds to the formation of surface dimers, which stabilized the electronic structure energy. Such a stabilization mechanism is commonly observed in covalent materials, as explained above. Except the creation of the 'pseudo gap', however, the total DOS is not significantly changed between the two snapshots, since the fracture occurs locally and the rest part of the system keep the character of the bulk electronic structure. A drastic change is seen in the partial DOS $D_i(\varepsilon)$ between the two snapshots. In the snapshot (a), the partial DOS is distributed among almost all the energy range of the occupied levels, which is a typical character of the bulk Wannier state. In an ideal diamond crystal, all the bond sites are symmetrically equivalent and the partial DOS of a Wannier state is proportional to the DOS of the valence band. In the snapshot (b), on the other hand, the partial DOS has the main sharp peak near the atomic sp^3 level ($\varepsilon_{\rm h}$=-0.4eV). The sharpness of the peak implies that this wave function should be similar to the eigen state of an sp^3 dangling bond orbital. The partial DOS has also a sharp peak at a low energy region near the atomic s level ($\varepsilon_{\rm s}$=-5.45eV). The contribution of such a low energy atomic level is essential for the dehybridization mechanism in the present elementary process, as discussed above.

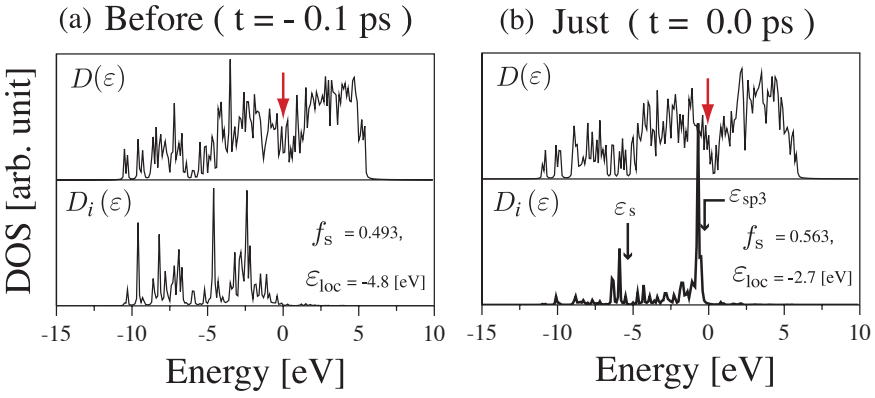


Figure 7.7: Total DOS $D(\varepsilon)$ and Partial DOS for the Wannier state $D_i(\varepsilon)$ in the course of the bond breaking process described in Fig. 7.5. The snapshot (a) is a snapshot before the bond breaking ($t = -0.1$ps) and (b) is that just on the bond breaking time ($t = 0$). The chemical potential is indicated as the red arrows.

Here we emphasis the crucial importance of the quantum mechanical freedoms in the above process. The importance can be seen in the fact that the bond breaking do *not* occur at a single bond site, but occur at two successive bond sites. The bond breaking at a *single* bond site is quite difficult to occur, because the resultant single dangling bond state would be quite instable. The bond breaking at successive *two* bond sites, on the other hand, is easy to occur, because the resultant two-fold coordinated atom can have a atomic lone pair state and can be stabilized by the dehybridization mechanism. The importance can be also seen in the fact that the peak value of ε_i, the value at $t = 0$ in Fig. 7.5 is not unique among the Wannier states in the fracture process. The peak value is affected by the other Wannier states $|\phi_j\rangle$ under the orthogonality constraint. The importance of the orthogonality between the Wannier states can be seen in the following fact of Fig.7.6(d); the atomic s level ($\varepsilon_{\rm s} = -5.45$eV) is lower than the energy level of the bulk Wannier state ($\varepsilon_{\rm WS} = -5.08$eV). This fact means that the bulk Wannier state $|\phi_i\rangle$ at $\varepsilon = \varepsilon_{\rm WS}$ would occupy immediately the atomic s orbital, *if it was vacant*. In summary, the elementary fracture process is described by the energy competition among several Wannier states and the essential quantum mechanical freedoms are the dehybridization mechanism and the orthogonality constraint.

7.3 Effect of dehybridization mechanism

From the analysis in the previous section (Section 7.2), we found that the dehybridization mechanism is essential for the elementary fracture process. In this section, we will clarify the importance of the dehybridization mechanism using the comparison with an artificial material.

Preparation of artificial material

The artificial material is given by a modified tight-binding Hamiltonian of silicon. The modification is done by tuning the difference of the atomic p and s levels to be zero ($\varepsilon_{\rm p} - \varepsilon_{\rm s} = 0$). The corresponding metallicity parameter is given by $\alpha_{\rm m} = 0$. All the other parameters in the tight-binding Hamiltonian are the same as that in the silicon case. We have discussed, in Chapter 3, that such a parameter tuning reproduces the variety among the group IV elements within the universal tight-binding theory. The artificial material is a semiconductor with sp^3 bonding states. The values of the band gap and the band width are comparable to those in silicon, as is seen below. The weight of s orbital is $f_{\rm s} = 0.247$ in the bulk state, which means an almost ideal sp^3 hybridization. The crucial difference of the artificial material from silicon is the lack of the dehybridization mechanism. Since the dehybridization mechanism originates from the energy gain due to the increase of s component, such mechanism can not be seen in the artificial material because of $\varepsilon_{\rm p} - \varepsilon_{\rm s} = 0$. We will see that the lack of the dehybridization mechanism is crucial for the simulation result.

Silicon case ('617A' sample)

Before the comparison, we discuss the silicon case. So as to eliminate the numerical error of the order-N method, the exact diagonalization is used here. The sample is cubic with 617 atoms, which is a larger sample than that in the previous section. Unlike in the previous section, the present sample does not contain the initial defect bond. The present sample is referred to '617A' sample. The results of the fracture simulation are shown in Fig. 7.8 and Fig. 7.9. The time $t = 0$ corresponds to the beginning of imposing the external load. The fracture begins at $t \approx 4.5$ps and the sample is divided into two peaces. The final structure will be shown later in this section. Figure 7.8 shows the eigen energies near the highest occupied or lowest unoccupied levels. Figure 7.9 shows the surface stress and the eigen energies as the function of time. Unlike the figures in the previous section, here we plot all the eigen levels in the lower panel of Fig. 7.9. At the initial structure, the highest occupied and lowest unoccupied levels locate at by $\varepsilon \approx 0$eV and $\varepsilon \approx 1.3$eV, respectively.

Now several comments are added; One may find another energy 'gap', in a lower energy region ($-8\text{eV} \leq \varepsilon \leq -7\text{eV}$). This 'gap', however, is due to the finite size effect and is not essential. One may observe a global property of the sample in Fig. 7.9; the band widths of the valence and band bands decrease with increasing the external load before the fracture ($t < 4.5$ps) but turn to increase after the fracture This is understandable as follows; In general, a band width is determined by the transfer integral. If a lattice is deformed with a longer lattice constant, the transfer integral and the band width will decrease.

Here we discuss the electronic structure in the present fracture simulation. Though the initial defect bond is not prepared in this sample, several 'defect' levels within the band gap can be found in the initial structure. They are separated from the continuum valence or conduction band. These 'defect states' should originate from the sample surfaces. When the fracture begins at the critical external load, on the order of GPa, the electronic energy gap vanishes. After the fracture, an electronic energy gap appears again with a smaller value than that of the initial structure. The energy gap of the final structure corresponds to that in the surface band. Since the microscopic pictures are the same in the previous section, we do not repeat them here.

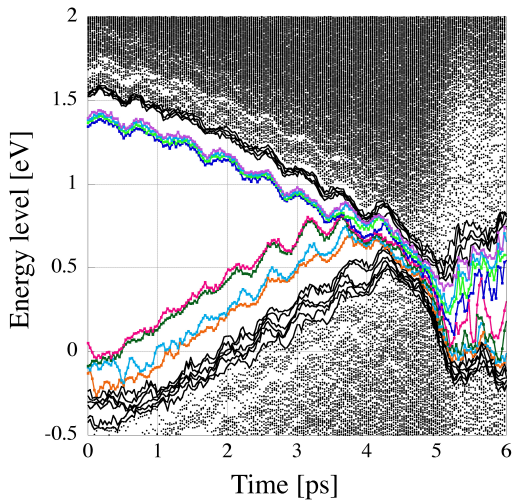


Figure 7.8: Selected eigen energies in the '617A' sample as the function of time. The highest occupied and the lowest unoccupied levels is plotted as the red and blue lines, respectively. Several levels near the highest occupied and lowest unoccupied levels are plotted as (colored) lines. The other levels are plotted as dots.

Comparison with artificial material ('617B' sample)

The fracture simulation of the artificial material, or the modified Hamiltonian, is done with the same sample geometry as the silicon case ('617A' sample). This sample of the modified Hamiltonian is referred to '617B' sample. The simulations of the '617A' (silicon) and '617B' (artificial material) samples are done with the same conditions except the modification of the tight-binding parameters. Figure 7.10 shows the resultant final structures of the silicon case (a) and the modified Hamiltonian case (b). Unlike the silicon case, the artificial material is not divided into two peaces, but shows a formation of a disordered or amorphous region. Figure 7.11 shows the dynamics of forming an amorphous region. The results in the '617B' sample are shown in Fig. 7.12 and Fig. 7.13, which are compared with those in the '617A' sample, Fig. 7.8 and Fig. 7.9, respectively. One may find a difference between the two materials, as the fact that the stress values at $t = 0$ in the '617B' sample is deviated from zero. This is because the equilibrium bond length is different from

the silicon case, which is not essential for the fracture behavior. Figure 7.13 shows the surface stress and the eigen energies, as the functions of time. The artificial material has a finite energy gap that is comparable to that of the silicon case. When the fracture begins, the surface stress is given on the order of GPa, and the electronic energy gap almost vanishes. In the final structure, an electronic energy gap appears but is smaller than that of the initial structure. All the above features in the eigen energy levels is quite similar to those in the silicon case, in Fig. 7.9.

Hereafter we will discuss why the two cases differ in the final structures, though they are similar in the stress and the eigen-value distribution. The creation of a cleaved surface and a local amorphous region can be understood by the dynamical structural changes so as to release the bulk strain energy. The releasing the bulk strain energy can be seen commonly in the lower panel of Fig. 7.9 and Fig. 7.13, as recovering the band width after the stress has its peak. The difference of the two cases is the presence or the absence of the stability mechanism of a cleaved surface. In the silicon case, as explained in Fig. 7.5, an two-fold coordinated atom appears after the bond breaking. The two-fold coordinated atom is stabilized by the dehybridization mechanism, or the increase of s components ($f_{\rm s}$). This dehybridization mechanism is essential, because the atomic level difference ($\varepsilon_{\rm p} - \varepsilon_{\rm s}$) is so large to be comparable with the bond breaking energy ($2|\beta_0|$). In the present artificial material, on the other hand, the atomic level difference is zero ($\varepsilon_{\rm p} - \varepsilon_{\rm s} = 0$) and an *unreconstructed* surface cannot be stabilized by the dehybridization mechanism.

The present comparison shows the essential role of the dehybridization mechanism in the present brittle fracture of silicon, which is seen as the fact that the brittle fracture does not occur in a system without the dehybridization mechanism.

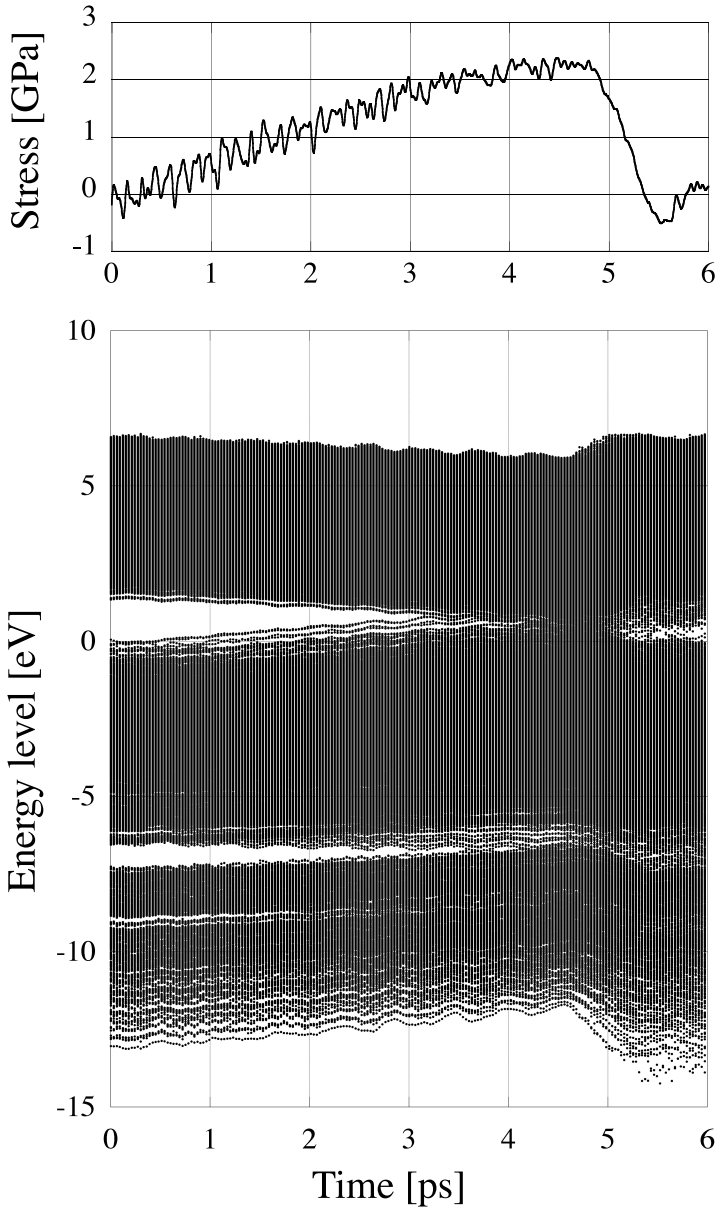
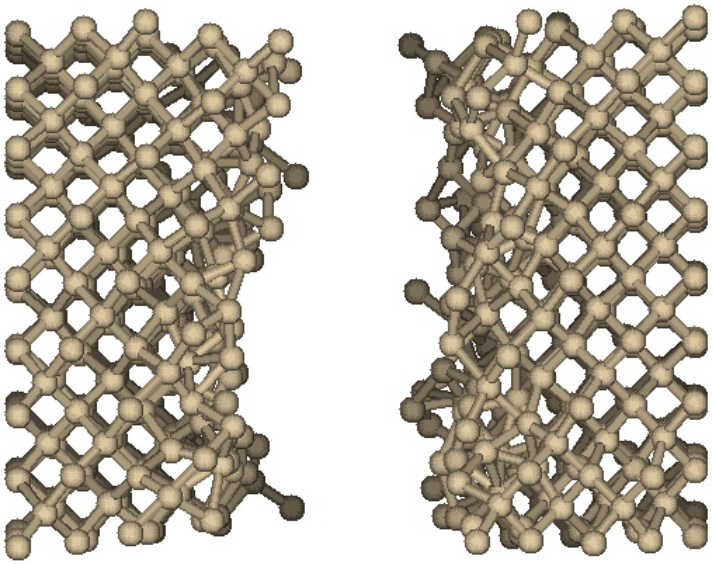


Figure 7.9: The stress and all the eigen energies in the '617A' sample as the function of time. All the energy levels are plotted as dots.

(a) Silicon case



(b) Modified Hamiltonian

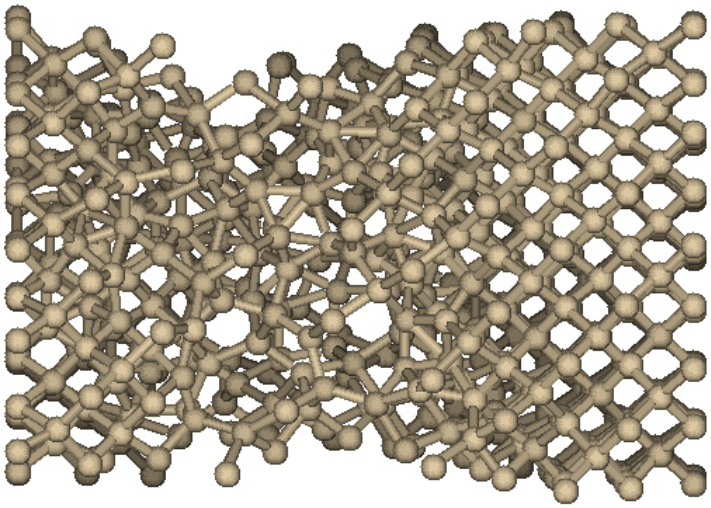


Figure 7.10: The final structures of (a) the silicon case ('617A' sample) and (b) the modified Hamiltonian ('617B' sample). Here the bonds are drawn, just for an eye guide, according to the interatomic distance.

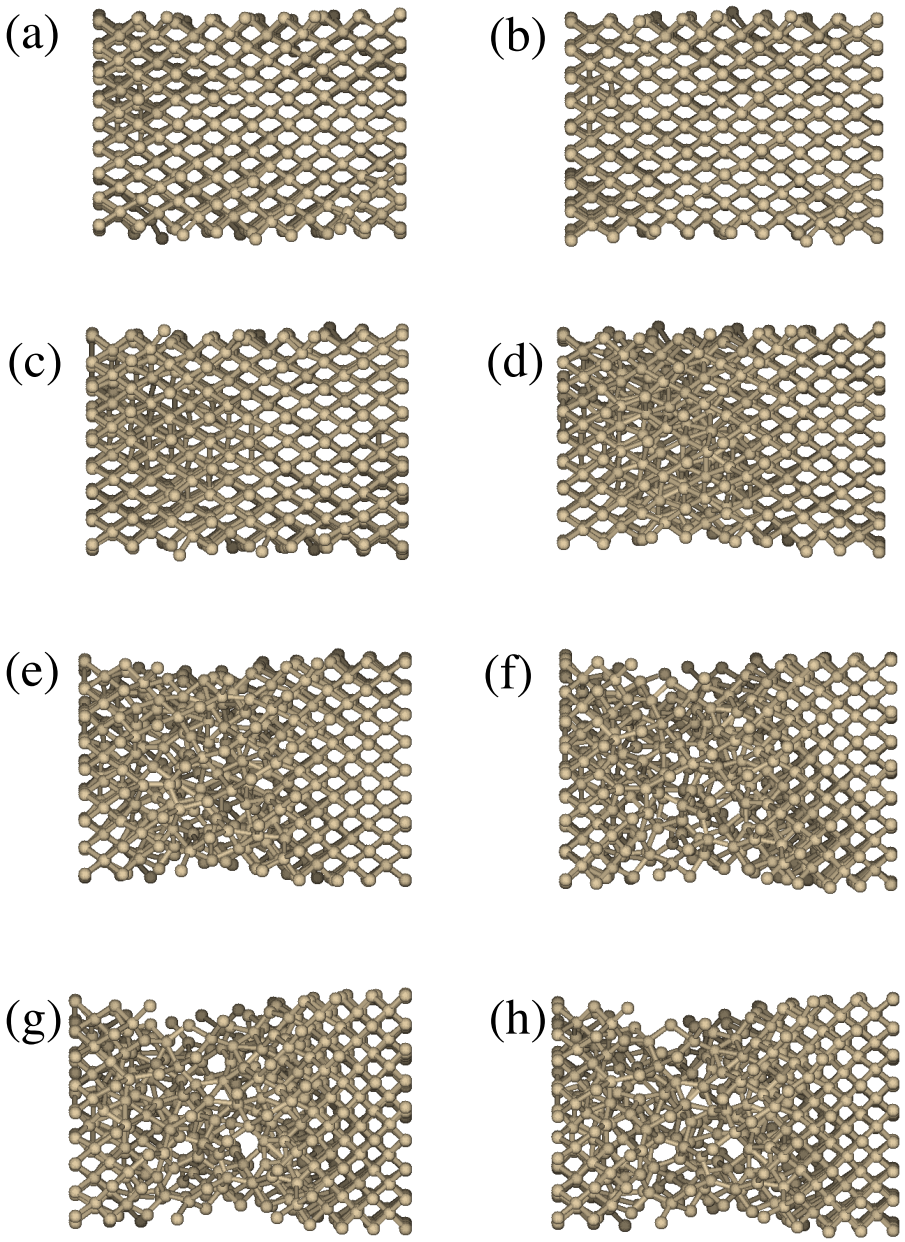


Figure 7.11: Snapshots of the dynamical fracture simulation of the '617B' sample with 617 Si atoms. The time interval between successive two snapshots is $\Delta t = 0.3$ ps. Here the bonds are drawn, just for an eye guide, according to the interatomic distance.

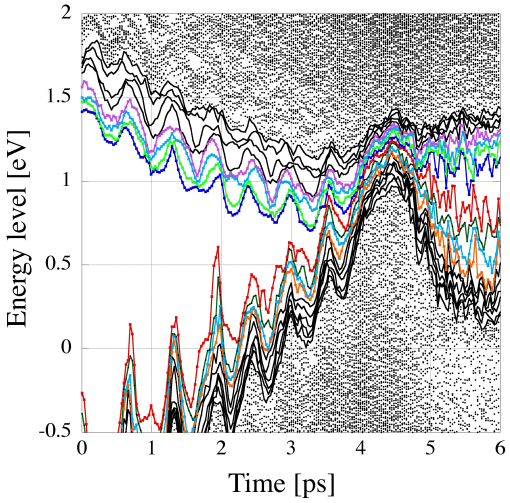


Figure 7.12: Selected eigen energies in the '617B' sample as the function of time. The plotting manner is the same as in Fig. 7.8.

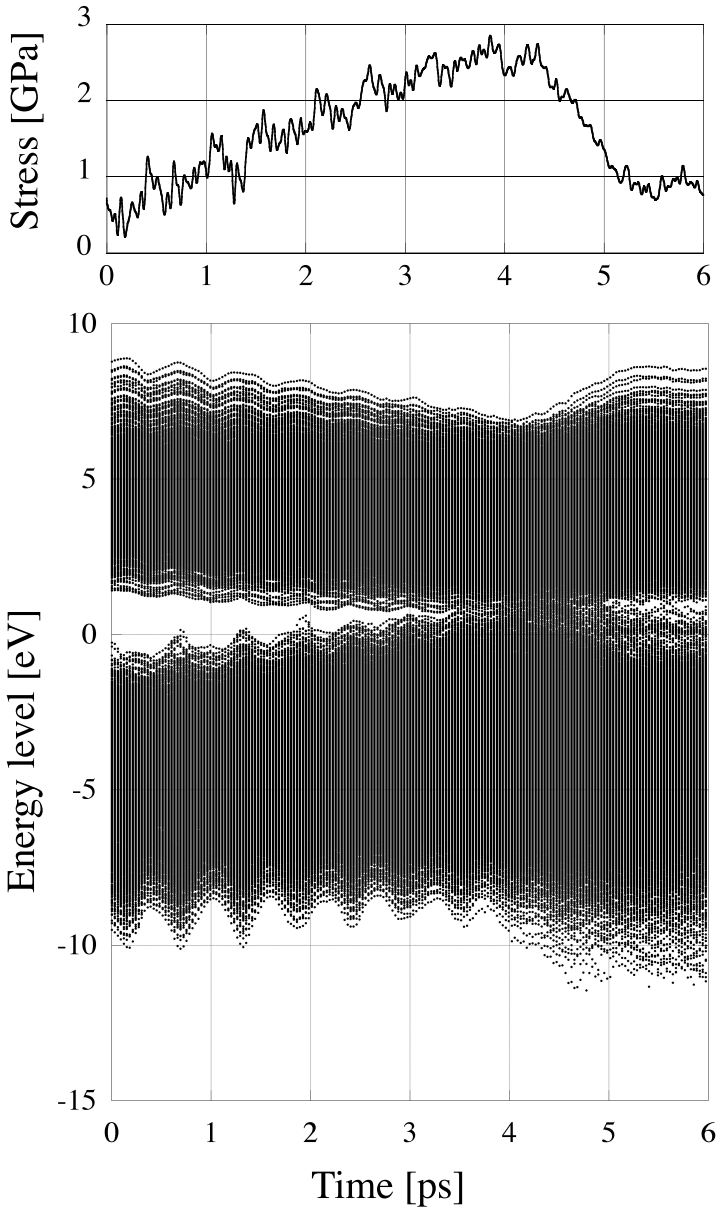


Figure 7.13: The stress and all the eigen energies in the '617B' sample as the function of time. The plotting manner is the same as in Fig. 7.9.

7.4 Technical details of the dynamical simulation

Here we discuss several technical details of the large-scale simulations that will be done in the rest of this chapter. The fundamental theory for large-scale simulations was discussed in Chapter 4 and 5. This section explains additional details, especially, details for dynamical simulations. Two points are mainly discussed.

Dynamical control of localization region for Wannier state

The first point is how the localization region for each Wannier state is determined in the variational order-N method, which governs the balance of the accuracy and the computational costs. As a test calculation, we calculate fracture simulations of a small system size (91 atoms) using different controlling methods. Figure 7.14, show the stress value, so as to monitor the accuracy of the simulation methods.

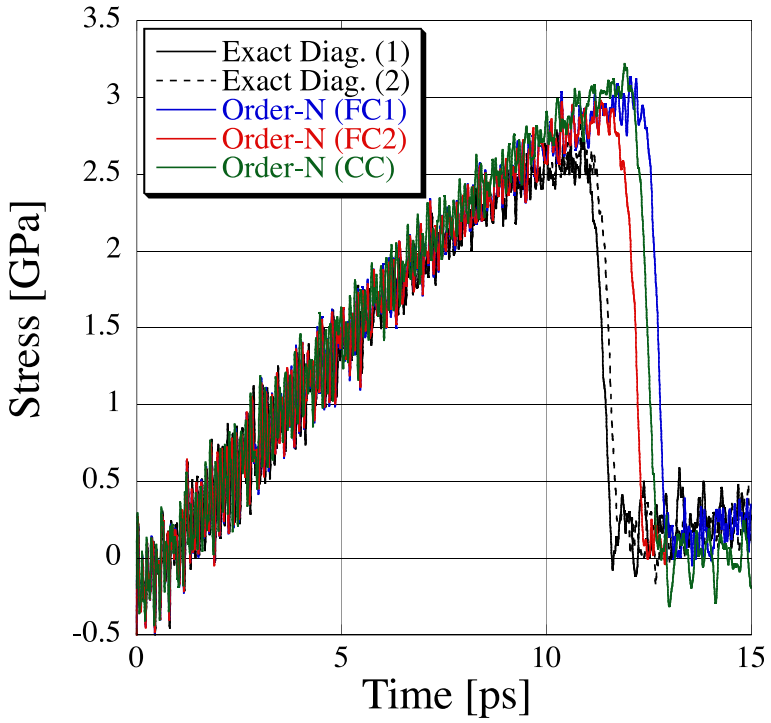


Figure 7.14: The stress value in the fracture simulations with different methods; Exact diagonalization methods with different temperature parameters in the Fermi-Dirac form are denoted as 'Exact Diag. (1)' and 'Exact Diag. (2)' . The variational order-N method with the constant cutoff radius for the localization constraint is denoted as 'Order-N (CC)'. The variational order-N methods with the flexible cutoff radius are denoted as 'Order-N (FC1)' and 'Order-N (FC2)', whose difference is discussed in the text.

We calculate the exact matrix diagonalization method as reference data. When the fracture occurs, the electronic energy gap vanishes, as seen in the previous sections. In the exact diagonalization method, the density matrix is given by the

fractional occupation with a finite temperature form

$$\hat{\rho} = \sum_k |\phi_k^{(\mathrm{eig})}\rangle \, f_k(\varepsilon_k) \, \langle \phi_k^{(\mathrm{eig})}|, \tag{7.5}$$

where the wave functions $\{\phi_k^{(\mathrm{eig})}\}$ are eigen states with eigen energies $\varepsilon_k^{(\mathrm{eig})}$. The occupation number $f_k(\varepsilon_k^{(\mathrm{eig})})$ is given by a Fermi-Dirac function with the chemical potential μ and a temperature parameter τ_{elec}

$$f_k(\varepsilon_k) = \frac{1}{1 + e^{(\varepsilon_k - \mu)/\tau_{\mathrm{elec}}}}. \tag{7.6}$$

The temperature parameter τ_{elec} is usually chosen for the numerical stability and is not necessarily equal to the temperature of the system. We achieve the exact diagonalization calculations with the different electronic temperatures; the case with $\tau_{\mathrm{elec}} = 0.1$[eV] is denoted as 'Exact Diag. (1)' in the figure, while that with $\tau_{\mathrm{elec}} = 0.01$[eV] is denoted as 'Exact Diag. (2)' .

We also calculate using the variational order-N method with different localization constraints. As explained in Section 4.3, we should control only the cutoff radius $r_i^{(\mathrm{cut})}$ for each Wannier state $|\phi_i\rangle$. We calculate with several controlling methods of the cutoff; the constant cutoff (CC), the flexible homogeneous cutoff (FC1) and the flexible inhomogeneous cutoff (FC2), which will be explained below. The resultant stress values are shown in Fig. 7.14 with the symbols '(CC)' , '(FC1)' and '(FC2)'.

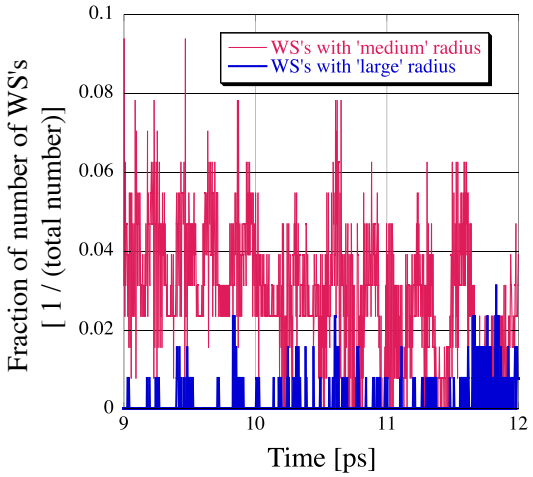


Figure 7.15: The fraction of the Wannier states with the 'middle' or 'large' cutoff radius. This quantity is used in the 'flexible inhomogeneous cutoff' method, which is denoted as '(FC2)' in Fig. 7.14.

Here we explain the above three cutoff methods in the variational order-N methods. In the initial sample without deformations, we choose the perturbative Wannier state as the initial state of the variational order-N method. Since the perturbative Wannier state extends over 20 atoms, as shown in Fig. 5.7, the cutoff radius should be, at least, enough large to contain these atoms. In the above case of the constant cutoff (CC), we choose the cutoff radius as the constant value of $r_i^{(\mathrm{cut})} = 2.5 d_0$, where $d_0 (= 2.35\text{Å})$ is the equilibrium bond length. This value is chosen for all

Wannier states through the simulation. Without an external load, this cutoff sets the localization region of the Wannier states to about $N_{\rm loc} = 40$ atoms. We should say, however, that the choice of the constant cutoff method is not appropriate in the present fracture simulation. In the present fracture simulation, the sample will be stretched by the external load. In other words, the system will be somewhat dilute and the number of atoms within a Wannier state will decrease, if a constant cutoff is used. A better way is to control the number of atoms in the localization region, not the radius in the distance $r_i^{(\rm cut)}$. Since the computational cost in the present tight-binding calculation is determined by the number of atoms, not the spatial volume, the computational cost will be not significantly changed with the above controlling method. We call such methods as 'flexible homogeneous cutoff' method, referred as '(FC1)' in Fig. 7.14. In this method, the cutoff radius $r_i^{(\rm cut)}$ is chosen so that the localization radius contains, at least, a finite number of atoms ($N_{\rm loc}^{(c)}$). We choose the value $N_{\rm loc}^{(c)} = 40$ in the '(FC1)' case in Fig. 7.14. The word 'homogeneous' is used, because the unique minimal number of atoms ($N_{\rm loc}^{(c)}$) is used among all Wannier states. In results, the cutoff radii $r_i^{(\rm cut)}$ may be different among the Wannier states but the number of atoms within the localization region ($N_{\rm loc}^{i}$) always satisfy $N_{\rm loc}^{(i)} \geq N_{\rm loc}^{(c)} = 40$. Now we explain the third method. In the present fracture simulation, the Wannier states only near the cleavage area are significantly changed. Especially at the beginning of the fracture, the electronic band gap is vanished and the localization constraint should be relaxed so as to reproduce the physical quantities. We can monitor the error ($|\delta\phi_i|$) of a Wannier state from its exact solution, as explained in Section 5.2. Using the above values, we can set *different* minimal atom numbers $N_{\rm loc}^{(c)}$ among different Wannier states. We call such methods as 'flexible inhomogeneous cutoff', which is denoted as '(FC2)' in Fig. 7.14. Here we prepare three minimal atom numbers, that is, $N_{\rm loc}^{(\rm c1)} \equiv 40$ for 'small radius' , $N_{\rm loc}^{(\rm c2)} \equiv 60$ for 'middle radius' , $N_{\rm loc}^{(\rm c3)} \equiv 80$ for 'large radius'. We also define the average of the errors among the Wannier states;

$$\delta\phi_{\rm av} \equiv \frac{1}{N}\sum_i |\delta\phi_i|. \tag{7.7}$$

If the error of a Wannier state ($|\delta\phi_i|$) is almost the same as its averaged value ($|\delta\phi_i| \leq 1.2\delta\phi_{\rm av}$), we require the minimal atom number for 'small radius' ($N_{\rm loc}^{(i)} = N_{\rm loc}^{(\rm c1)} \equiv 40$) to the Wannier state . If the error of a Wannier state ($|\delta\phi_i|$) is slightly larger than its averaged value ($1.2\delta\phi_{\rm av} \leq |\delta\phi_i| \leq 1.5\delta\phi_{\rm av}$), we require the minimal atom number for 'middle radius' ($N_{\rm loc}^{(i)} = N_{\rm loc}^{(\rm c2)} \equiv 60$) to the Wannier state. If the error of a Wannier state ($|\delta\phi_i|$) is larger than 150 % of its averaged value ($1.5\delta\phi_{\rm av} \leq |\delta\phi_i|$), we require the minimal atom number for 'large radius' ($N_{\rm loc}^{(i)} = N_{\rm loc}^{(\rm c3)} \equiv 80$) to the Wannier state. Fig. (7.15) shows the fraction of the number of Wannier states that are required the minimal atom numbers for 'middle' or 'large' radii. We plot the period 9ps $< t <$ 12ps, in which the fracture begins. From the figure, we observe that the 'middle' cutoff ($N_{\rm loc}^{(\rm c2)}$) is required to less than 10 % among all Wannier states and the 'large' cutoff ($N_{\rm loc}^{(\rm c3)}$) is required to less than 3 % among all Wannier states. The increase of the computational cost due to the requirement of the 'middle' or 'large' radius is not serious in the total cost.

Here we compare the resultant stress by these five methods, shown in Fig. 7.14.

In Section 7.2, we derived three important physical quantities; (i) the Young modulus, from the gradient in the linear (small load) region, (ii) the critical stress for fracture, from the peak value, (iii) the fracture propagation speed, from the beginning and the end of the fracture. If one estimate these values from the graphs of the different methods, no significant difference is obtained. Though the above agreement is satisfactory for the discussion of the fracture mechanism, we examine here the difference extensively. We replot Fig. 7.14 with finer scales, which are shown in Figs.7.16 and 7.17. Figure 7.16 shows the stress within a region of small external load, Here we can see that, if we delete the data of the constant cutoff (CC) method, in (b), the trajectories among the other four methods will be in a better agreement. In other words, the flexible cutoff methods, denoted 'FC(1)' and 'FC(2)', are better controlling methods than the constant cutoff method. Moreover we can see in Fig.7.16(b) that the two flexible cutoff methods are indistinguishable in the trajectories. The two diagonalization methods are also indistinguishable in the trajectories. Figure 7.17. shows the stress in a region of small external load. We compare the four method except the constant cutoff method. We can see that no significant difference is found in the trajectories of the two diagonalization methods, as in Fig.7.16(b). The difference of the critical external load or the time for the beginning of fracture is due to the local fluctuation, which is not essential for the fracture theory. The same conclusion can be applied to the two order-N methods.

In summary, we have tested three controlling methods for localization constraints in the variational order-N method; the constant cutoff (CC) method, the homogeneous flexible cutoff (FC1) method and the inhomogeneous flexible cutoff (FC2) method. The latter two methods are slightly better than the former one, though it is not significant in the results of the present fracture simulation. Since this test is only one example, we do not make a general conclusion for the controlling methods. The important point is that we provide the practical methods for controlling the accuracy and the computational costs.

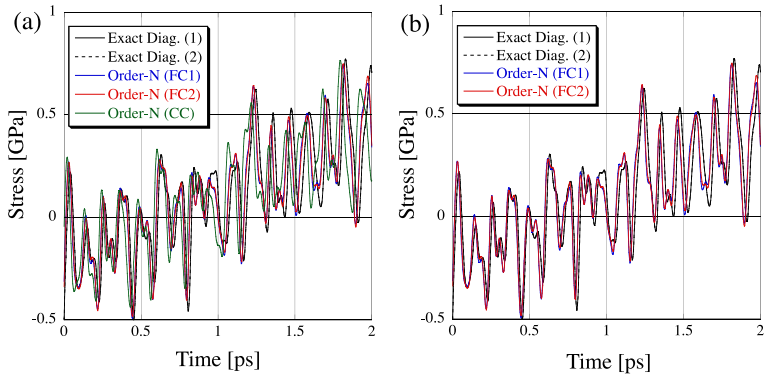


Figure 7.16: The stress value in the fracture simulations with different methods; All the data are the same as in Fig. 7.14 but are plotted with different scales on the axes. The two graphs are plotted with (a) and without (b) the constant cutoff (CC) method.

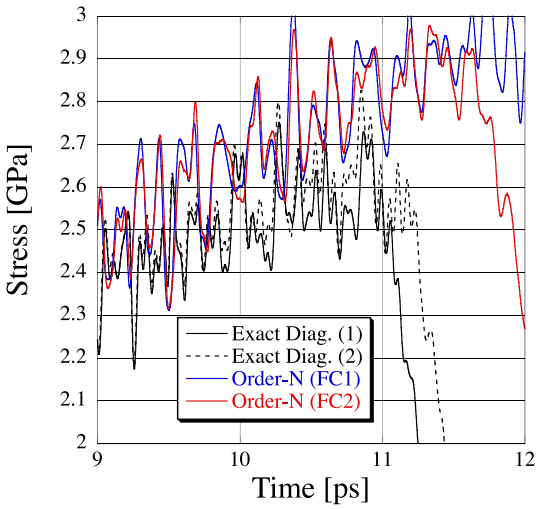


Figure 7.17: The stress value in the fracture simulations with different methods; All the data are the same as in Fig. 7.14 but are plotted with different scales on the axes. The graph is plotted without the constant cutoff (CC) method.

Dynamically-controlled hybrid scheme

The second point is the hybrid scheme in the dynamical simulation. In the fracture simulations with 10^4 atoms or more (Section 7.6), we use the hybrid scheme between the variational and perturbative order-N methods. In systems smaller than the above size (Section 7.5), the variational procedure is used in the whole region. As explained in section 4.4, the hybrid scheme works well by dividing Hilbert space. One technical point in the practical molecular dynamics is how to define the subsystem or how to select the member of the Wannier states treated in the variational method. In the fracture simulation, the variational method should be used only for selected Wannier states near fracture regions. During the fracture simulation, some of the variational Wannier states change their character dynamically from the bulk (sp^3 bonding) states to surface ones, as discussed before. The other wave functions, in bulk regions, keep the character of the bulk bonding state and can be obtained by the perturbative method.

In our program code, the member of the selected Wannier states can be dynamically controlled. We call the method 'dynamically-controlled hybrid scheme'. In the dynamical scheme, the treatment of one Wannier state can be switched from the perturbative method into the variational method. This switching is a 'one-way' algorithm. The present program code does not contain the 'reverse' switching, the switching from the variational method into the perturbative method. The 'reverse' switching will be discussed later. Now we restrict the switching to the switching from the perturbative method into the variational method. The switching is simply done by setting the perturbative wave function as the initial wave function for the iterative procedure in the variational method. Such a switching means the redefinition of the subsystems $\rho_{\rm A}$ and $\rho_{\rm B} \equiv \rho - \rho_{\rm A}$, which is well-defined in quantum mechanics and is automatically done in the program code. The algorithm is controlled by the criterion of the switching. The switching of a Wannier state ϕ_i should be done, when the perturbative treatment is broken down. A good quantity for the criteria is the weight of the non perturbative term, $|C^0|^2$ in Eq. (5.24). This value should be nearly equal to one ($|C_0|^2 \approx 1$) for the justification of the perturbative treatment. In the bulk crystal, the value of $|C^0|^2$ is $|C^0|^2 = 0.94$. Here we denote the weight as $w_i \equiv |C^0|^2$ for the i-th Wannier state ϕ_i. During the simulation with an external load, the value will decrease, due to the deformations. For the switching into the variational treatment, we can set a critical value $w_{\rm c}$ for the weight w_i as the lower limit for the perturbative treatment ($w_i > w_{\rm c}$). In general, an universal value for the criteria is not a good choice, because the value in the bulk case, $w_i = 0.94$ in the present silicon case, varies among the materials. A good reference for the critical value is the averaged value $w_{\rm av}$ among the weights of the perturbative Wannier states $\{w_i\}_i$. The critical value can be set, for example, as $w_{\rm c} = 0.95 w_{\rm av}$. Such a choice is reasonable, because the variational treatment should be done on the Wannier states in a local region with large deformations. The fracture will begin at such a local region.

Figure 7.18 shows an example of the dynamically-controlled hybrid scheme, in which selected Wannier states are plotted as atomic pairs of red balls. The sample contains 4501 Si atoms and one initial defect bond. The external load is imposed, but the fracture has not yet occurred. The figure is plotted in the following manner;

In the initial structure, or in the ideal crystalline geometry, all the Wannier states are treated in the perturbative method. The corresponding figure should contain no red ball. During the simulation with the [001] external load, a pair of red balls should appear in the figure, if one Wannier state is switched into the variational treatment. In Fig. 7.18, we observe a region of many red balls inside the sample, which is a region near the initial defect bond. This is because a large deformation is introduced in this region, due to the presence of the initial defect bond. The corresponding Wannier states are fairly deviated from bulk ones. We also observe a region of many red balls near the sample edges. This is because the corresponding Wannier states, those near the sample edges, are deviated from bulk ones, due to the difference of the environment. The above selection of the variationally treated Wannier states is quite reasonable, because the regions near the initial defect bond and near the sample edges are the candidates for the fracture seed. We will see the fracture simulation of this sample, in Section 7.5, using the variational method for all Wannier states. In results, the fracture begins at the initial defect bond, if it is contained. If not, the fracture begins at bond sites at the sample edge. In short, the dynamically-controlled hybrid scheme is done with the switching from the perturbative treatment into the variational one. The switching procedure is done automatically with a reliable threshold for each Wannier state.

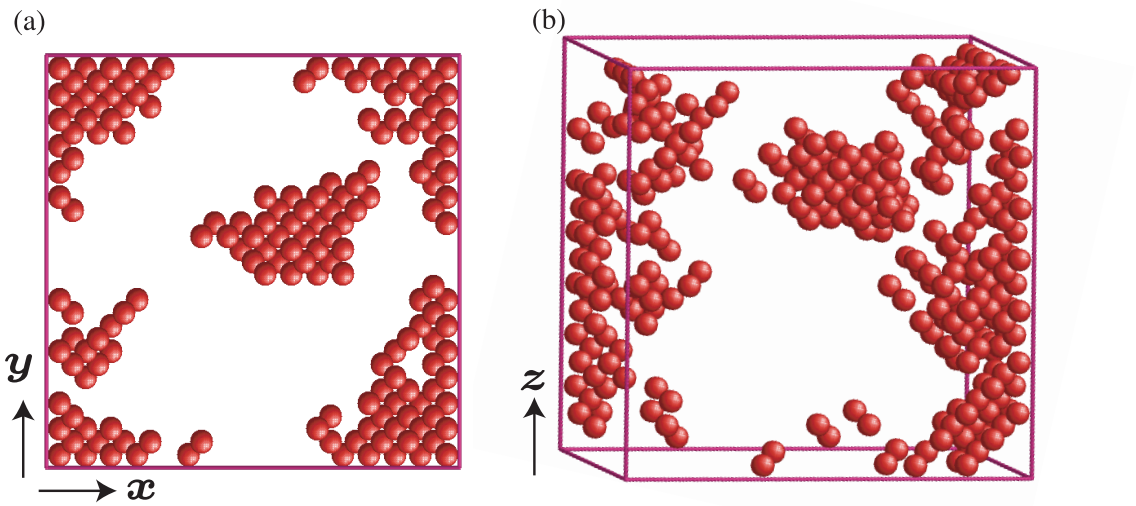


Figure 7.18: Top (a) and three-dimensional (b) views of selected Wannier states for the variational treatment in the dynamically-controlled hybrid scheme. The sample contains 4501 atoms and the sample edges are plotted as lines. A selected Wannier states is plotted as an atomic pair of red balls. Other atoms are invisible. The figures are plotted in the ideal crystalline geometry for an eye guide, though the actual sample is deformed with an external load in the [001] direction.

Though the dynamically-controlled hybrid scheme works fine in the above example, we do not use the scheme in the practical fracture simulations in this thesis, One practical problem for large-scale simulations is the limitation of the built-in memory size. In the variational method, unlike the perturbative method, the wave

function of the Wannier state should be stored in the memory. If one Wannier state is switched into the variational treatment, the required memory size will increase so as to store the Wannier states. In fracture simulation, as explained above, the number of the variationally treated Wannier states will increase with fracture propagation. If the built-in memory size is not enough large, the simulation has a possibility to exceed the limitation of the built-in memory size. Of course, the possibility depends on the balance of the sample size and the hardware environment. The dynamically-controlled hybrid scheme will be used in different systems and/or different hardwares. Another problem in the present dynamically-controlled hybrid scheme is the lack of the 'reverse' switching, the switching from the variational method into the perturbative method. The 'reverse' switching may be important, for example, in the simulation of crystal growth, because the number of Wannier states in the bulk (sp^3) bonding character will increase due to the crystal growth. It is also noteworthy that the switching from the variational method into the perturbative method will save the required memory size. A practical algorithm for the 'reverse' switching is one of future works.

Other technical details

Finally, we explain several tips that are used in the present hybrid scheme for better numerical results; One is the correction of the absolute value of the total energy. The one-electron energies of the perturbative wave function $\phi_i^{\rm PT}$ are slightly different from those of the variational wave function $\phi_i^{\rm VR}$. Its typical values is less than 0.1 eV, as explained in Section 5.5. The above difference

$$\Delta\varepsilon_i \equiv \langle\phi_i^{\rm PT}|H|\phi_i^{\rm PT}\rangle - \langle\phi_i^{\rm VR}|H|\phi_i^{\rm VR}\rangle \tag{7.8}$$

may cause a discontinuity of the absolute value of the total energy, when the wave function is switched from the perturbative treatment into the variational one. To avoid the discontinuity, we correct the absolute value of the total energy $\Delta\varepsilon_i$, when the switching occurs. This correction is made only for the absolute value of the total energy, which does not affect any atomic or electronic structure. Another tip is the correction of the equilibrium bond length. As discussed in Section 4.3, the equilibrium bond length in the perturbative method is slightly deviated, by about 2 % from the correct value. If the hybrid scheme, the above difference may cause an artificial lattice mismatching between the regions of variational and perturbative Wannier states. We correct this problem by introducing an additional (classical) potential on the atomic pairs that are the centers of the perturbative Wannier states. If the switching occurs for one Wannier state, the corresponding additional potential will be vanished, which is automatically done in the program code. The above correction is reasonable, though no detailed investigation is made for the actual effect on the present fracture theory. From a general viewpoint, the lattice mismatching of 2 % is not negligible in some cases. For example, the lattice constants between silicon and germanium differ only by 4 %. The above difference is crucial, when one discuss the Si/Ge interfaces.

7.5 Fracture simulation with thousands of atoms

In the present and next sections, we present and analyze fracture simulations with larger samples. We will focus on the reconstructed structures on the cleaved surface, which can not be discussed in smaller samples. This section describe the sample with 4501 atoms. The sample is a cubic cluster that is labeled, by the number of atomic layers, as $n_{\rm x} \times n_{\rm y} \times n_{\rm z} = 33 \times 33 \times 33$. The sample size in the length scale is given as $L_{\rm x} \times L_{\rm y} \times L_{\rm z} = 4.344\,{\rm nm} \times 4.344\,{\rm nm} \times 4.344\,{\rm nm}$. Simulations are done by the following conditions. Within the atoms on the 'top' or 'bottom' sample surfaces, the z component of their motion is under the constraint of the constant-velocity motion, as in Fig. 7.3. No other constraint is imposed on atomic motions. Two samples with 4501 atoms are simulated. One is the sample *with* the initial defect bond, which is denoted as '4501A' sample. The other is the sample *without* the initial defect bond, which is denoted as '4501B' sample.

Here a quantum mechanical analysis will be done in the simulation results. As discussed in Section 7.2, the elementary fracture process accompanies the two-stage reconstruction process of Wannier states. To observe such elementary processes, all the Wannier states are classified into bonding orbitals or atomic orbitals, according to the weight distribution among atoms. The bonding orbital is shown as a rod, while the atomic orbital is shown as a ball. Moreover, we use the color for the further analysis. For bonding states, the black rods are the reconstructed bonds that are not seen in the initial (crystalline) structure, while the white rods are the bulk bonds that are seen in the initial structure. For atomic states, the colors of the ball correspond to the weight of the s orbitals ($f_{\rm s}^{(i)}$); (i)$0 \le f_{\rm s}^{(i)} \le 0.2$ for the blue ball, (ii)$0.2 \le f_{\rm s}^{(i)} \le 0.3$ for the cyan ball, (iii)$0.3 \le f_{\rm s}^{(i)} \le 0.4$ for the white ball, (iv)$0.4 \le f_{\rm s}^{(i)} \le 0.5$ for the green ball, (v)$0.5 \le f_{\rm s}^{(i)} \le 0.6$ for the yellow ball and (vi)$0.6 \le f_{\rm s}^{(i)} \le 1$ for the red ball, respectively. Note that the bulk bonding state gives the value of $f_{\rm s}^{(i)} = 0.36$.

Sample the initial defect bond ('4501A'sample)

Figures 7.19, 7.20, 7.21, 7.22 show the result of the '4501A' sample, the sample with the initial defect bond. In the final structure, the sample contains a (001) cleavage plane atomistically flat except the area near the sample surfaces. We also observe, in Figs.7.21 and 7.22, that the atomic layer with the initial defect bond has grown as the cleavage plane.

The resultant cleavage surface contains many asymmetric surface dimers that have red or yellow balls, which corresponds to Fig. 3.2 in Section 3.2. Such a ball corresponds to the atomic 'π' state localized on the 'upper' atom with a high value of $f_{\rm s}$, as explained in Section 7.2. The resultant cleavage surface also contains many two-fold coordinated atoms that have two back bonds (white rods) and a lone pair atomic state (ball). This is because the two-fold coordinated atoms are metastable, as explained in Section 7.2. Since the present fracture simulation is a thermally non-equilibrium dynamics within a few pico seconds, such a metastable atom can be preserved until the end of the simulation.

Sample without the initial defect bond ('4501B'sample)

Figures 7.23, 7.24, 7.25, 7.26 show the result of the '4501B' sample, the sample *without* the initial defect bond. As we see in Fig. 7.23, the fractured sample contains two cleavage planes. In Fig.7.25, we observe, as in the '4501A' sample, the formation of an atomistically flat (001) surface with many asymmetric dimers except the areas near the sample boundary surfaces. Figures 7.25, 7.26 show the fracture dynamics.

As a crucial difference from the sample with the initial defect bond (the '4501A' sample), the fracture begins at two points on the sample edges in the present sample (the '4501B' sample). This is reasonable, when we think, as a general tendency, that a sample edge region should be mechanically weaker than a bulk region. Within the present simulation, the above tendency can be understood as follows; As discussed in Section 5.5, the Wannier states are stabilized by the transfer energy due to its spatial extension. For the bulk state, the corresponding energy gain is given by $\varepsilon_{\rm b} - \varepsilon_{\rm WS} \approx 0.6$ eV from Fig. 5.10. A bonding Wannier state near the sample edges does not have all the neighboring bond sites that the Wannier states in the bulk region have. Such a Wannier state has a smaller gain of the transfer energy than that in the bulk region. In results, a bond near the sample edges is weaker than a bond in the bulk region.

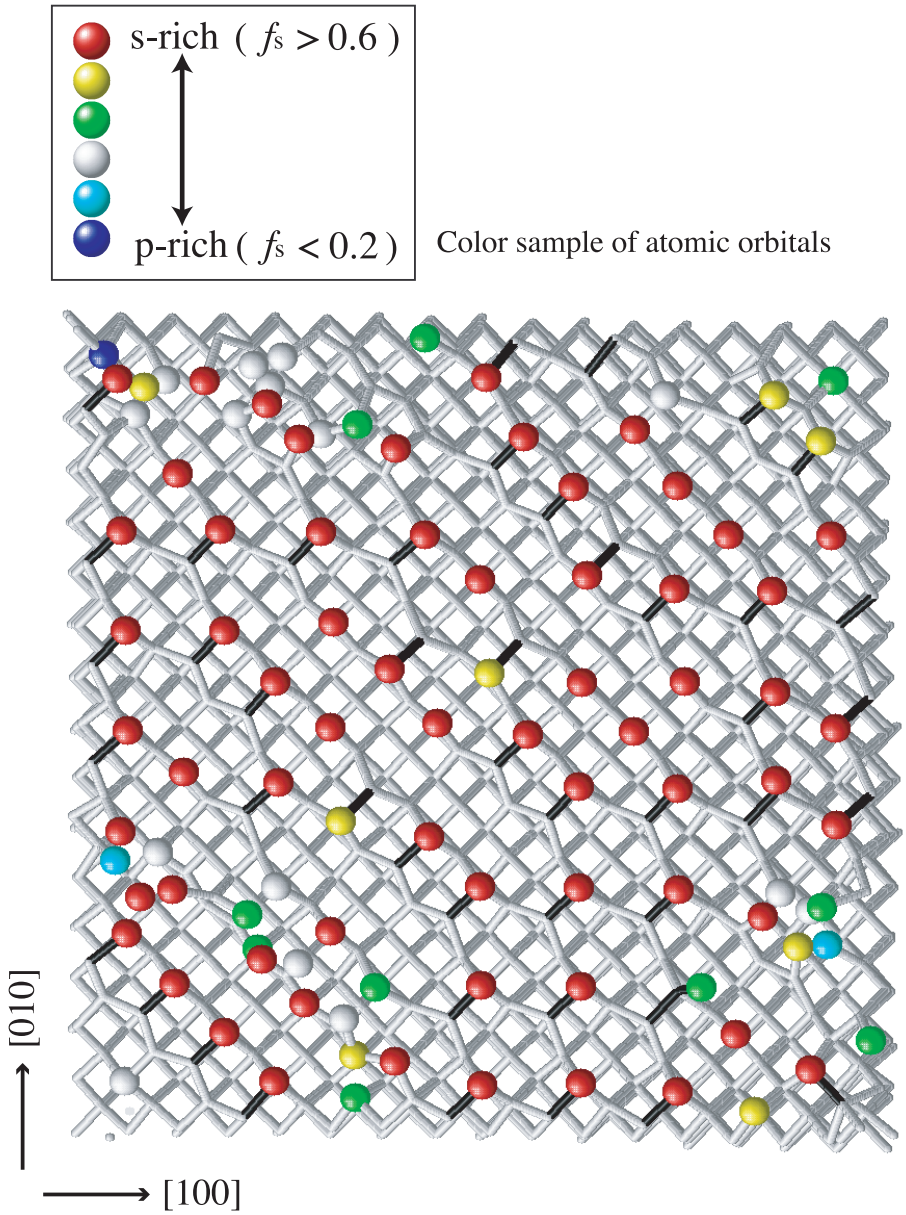


Figure 7.19: Top view of the cleaved plane in the fractured '4501A' sample with 4501 Si atoms. The black rods indicate the *reconstructed* bonding Wannier states, while the white rods indicate the bonding Wannier states that are contained in the initial crystalline structure. The atomic Wannier states are plotted as colored balls. The color of balls is determined by the weight of s orbitals (f_s). See the text for details.

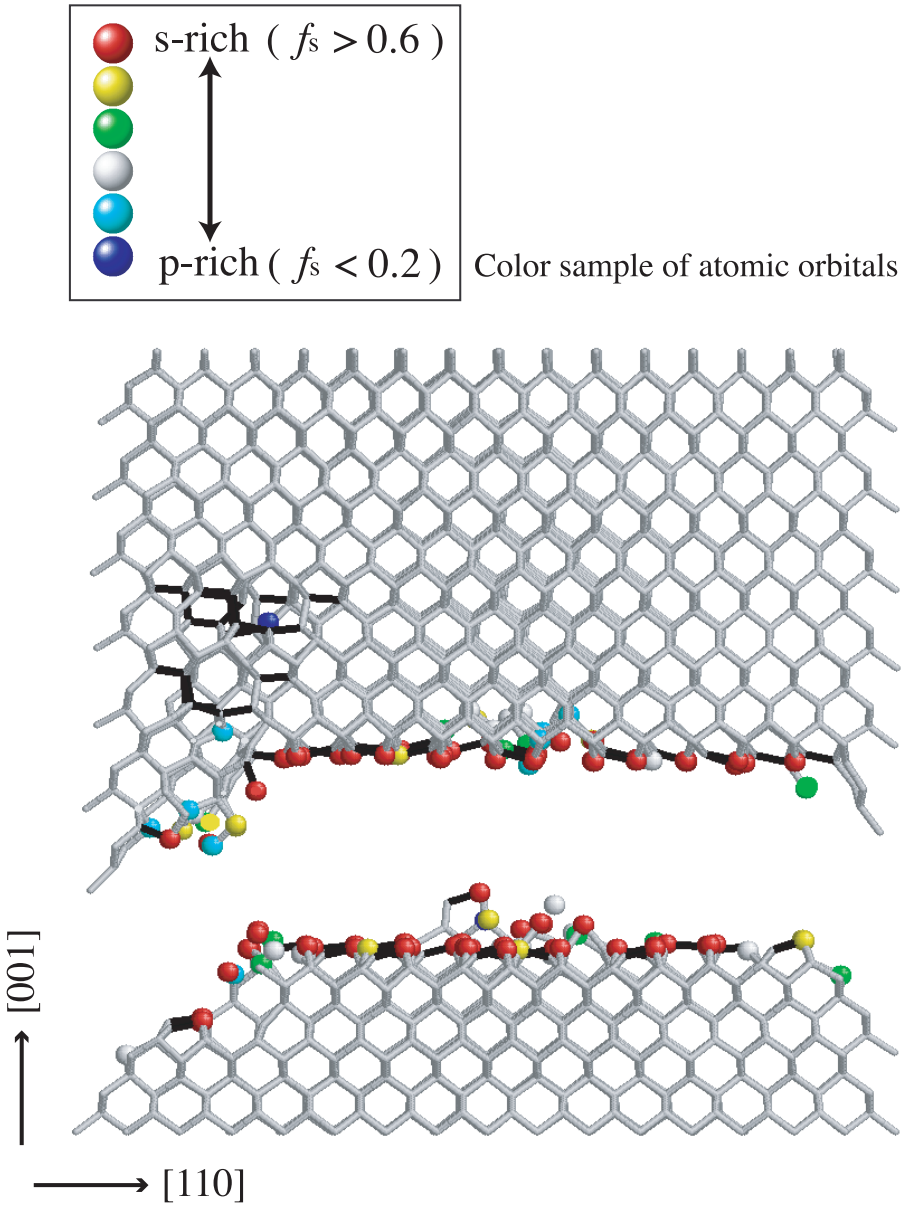


Figure 7.20: A 'semi-infinite' view of the fractured '4501A' sample with 4501 Si atoms. Only the atoms in the semi-infinite region ($y \leq x$) are visible and the other atoms are invisible. The present figure shows the same snapshot as in Fig. 7.19 but is plotted in the different view method. See the caption of Fig. 7.19 for the descriptions of the rods and the balls.

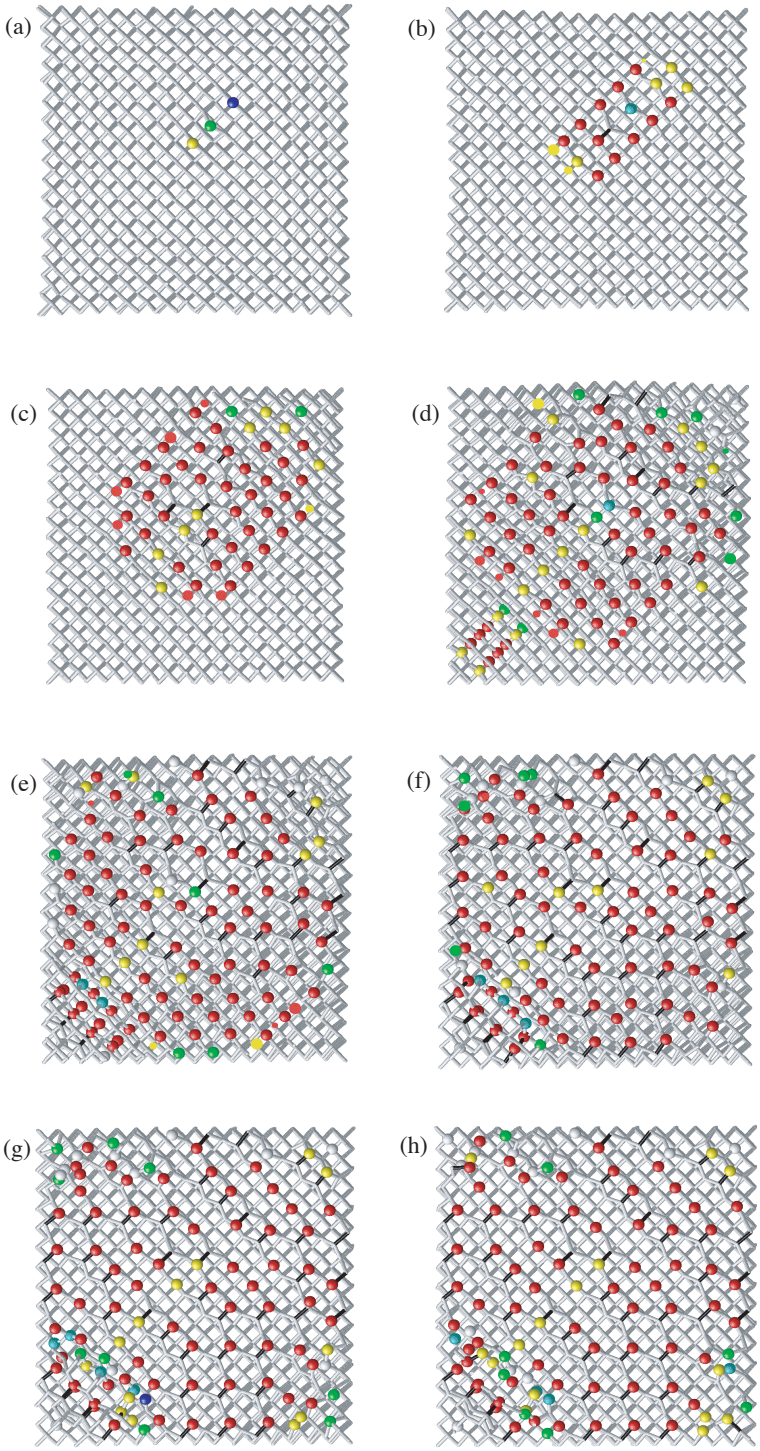


Figure 7.21: Snapshots of the dynamical fracture simulation of the '4501A' sample with 4501 Si atoms. The present figures are plotted in the same view method as in Fig. 7.19. The time interval between successive two snapshots is $\Delta t = 0.18$ ps. Figure 7.19 corresponds to the snapshot with a delay of 1.74 ps from the snapshot (h).

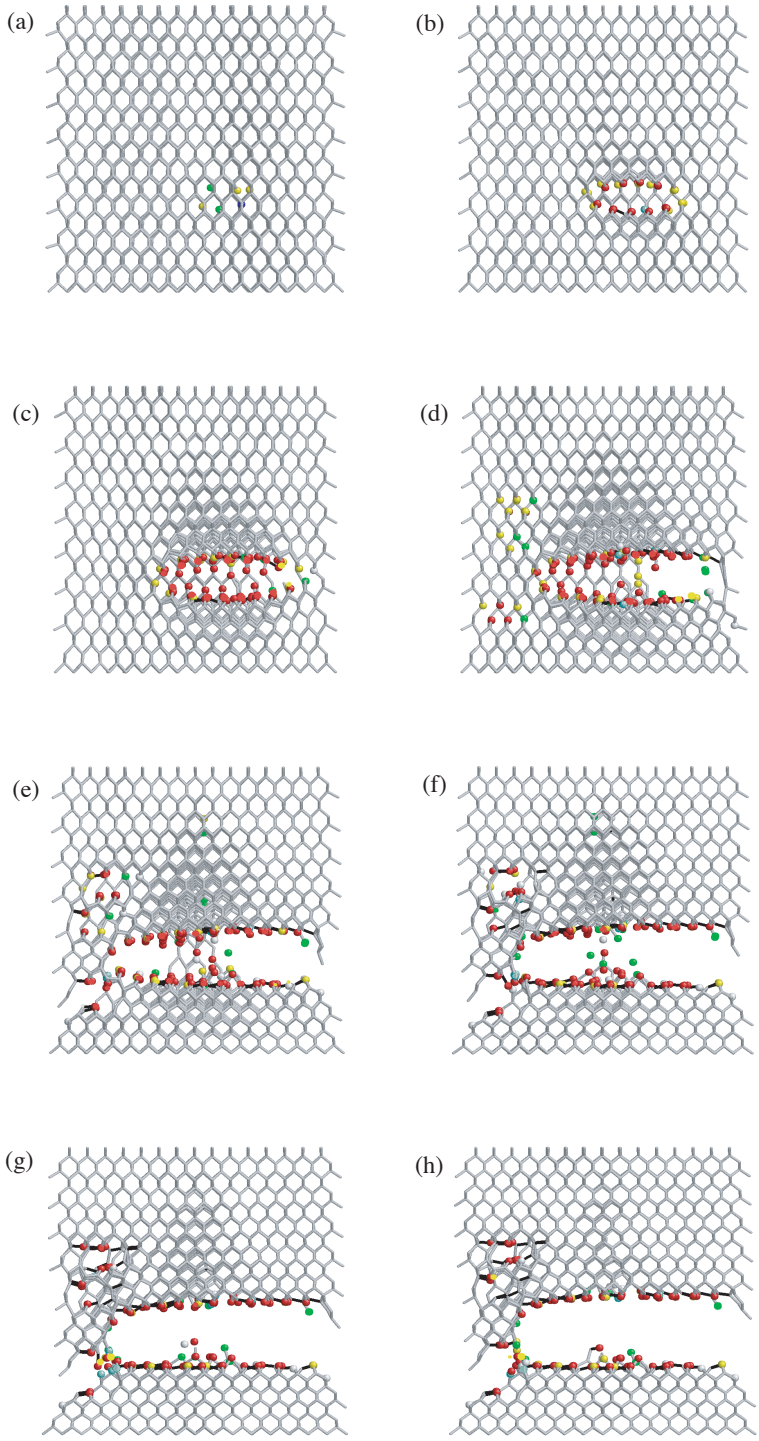


Figure 7.22: Snapshots of the dynamical fracture simulation of the '4501A' sample with 4501 Si atoms. The present figures (a) ,(b)... (h) show, in the different view method, the same snapshots Fig. 7.21 (a) ,(b)... (h), respectively. The present view method is described in the caption of Fig. 7.20.

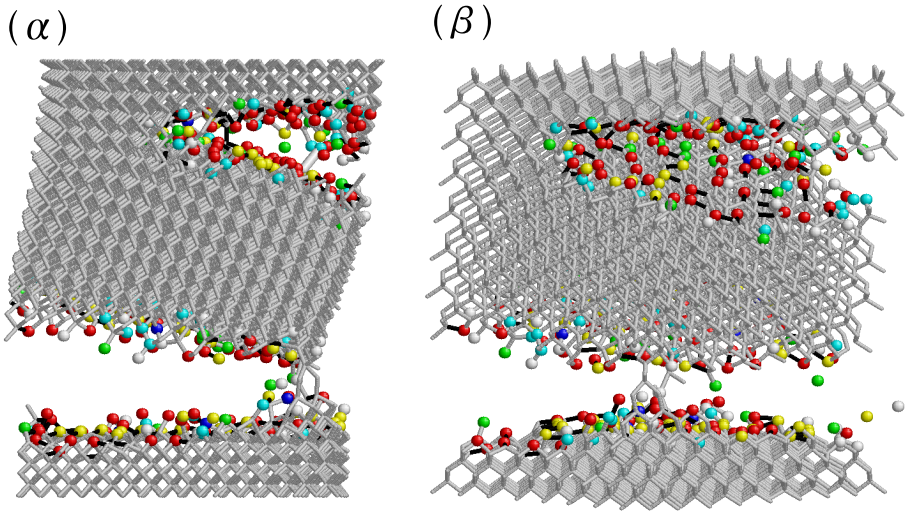


Figure 7.23: 3D views of the fractured '4501B' sample with 4501 Si atoms. See the caption of Fig. 7.19 for the descriptions of the rods and the balls. The figures (α) and (β) shows the same sample from different view points.

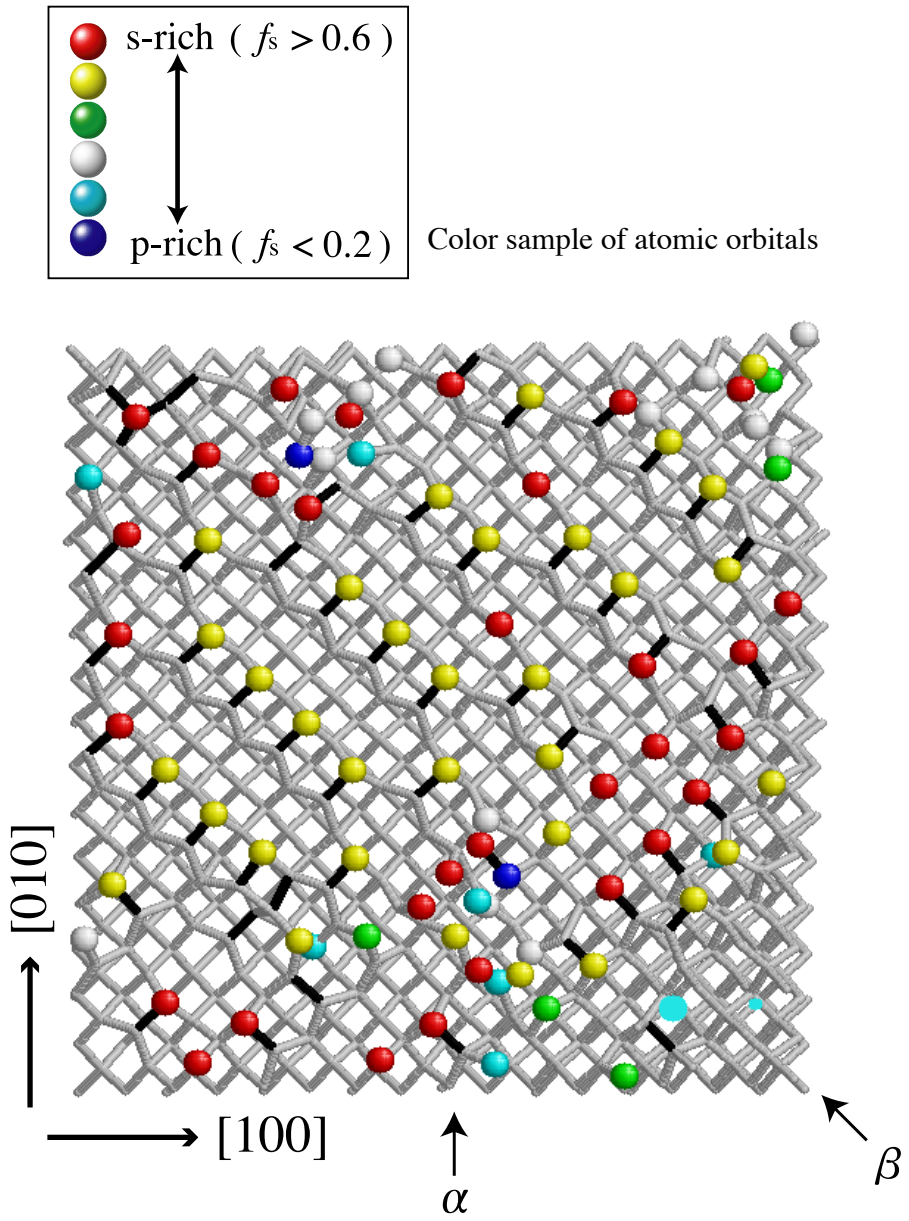


Figure 7.24: Top view of the cleaved plane in the fractured '4501B' sample with 4501 Si atoms. This plane corresponds to the lower cleavage plane in Fig. 7.23. The right-down corner has not yet fractured. See the caption of Fig. 7.19 for the descriptions of the rods and the balls. The arrows with 'α' and 'β' indicates the view directions in Fig.7.23 (α) and (β), respectively.

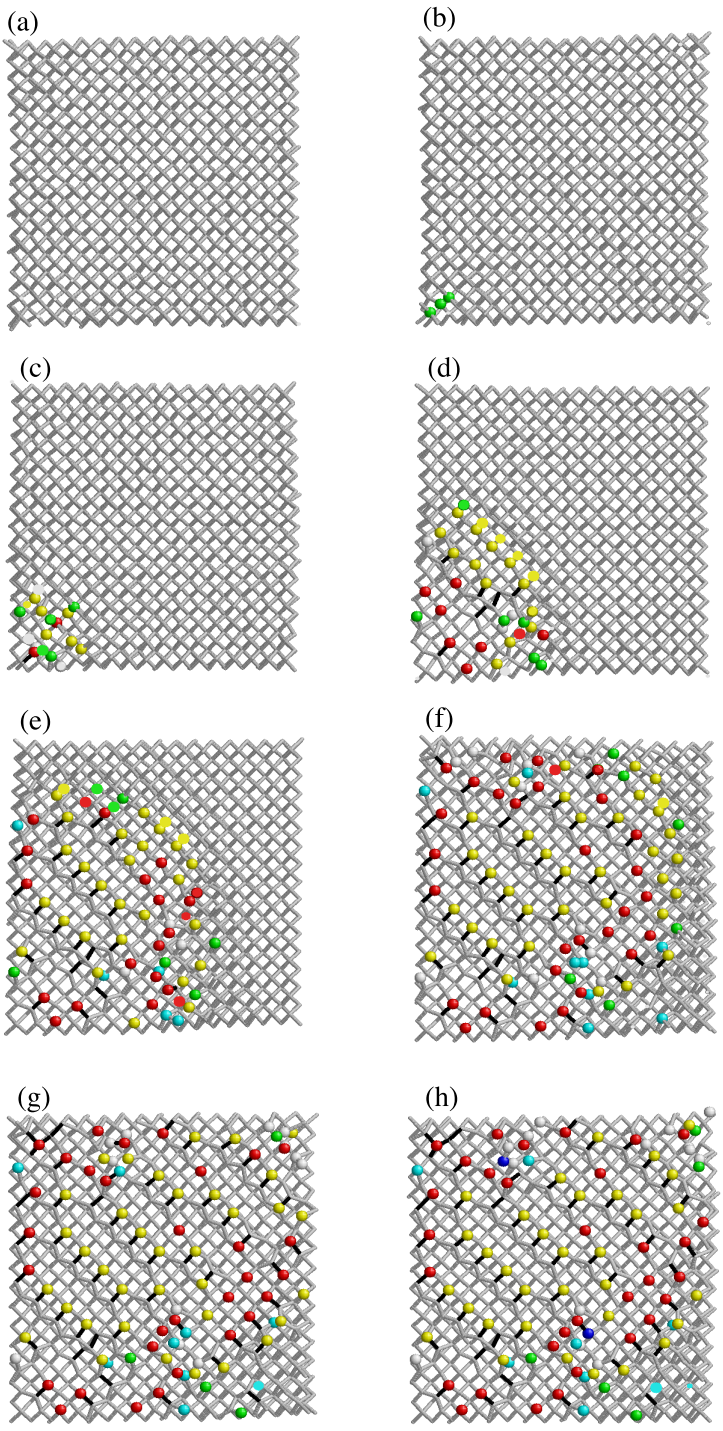


Figure 7.25: Snapshots of the dynamical fracture simulation of the '4501B' sample with 4501 Si atoms. The present figures are plotted in the same view method as in Fig. 7.24. The time interval between successive two snapshots is $\Delta t = 0.375$ ps. Figure 7.24 corresponds to the snapshot with a delay of 0.375 ps from the snapshot (h).

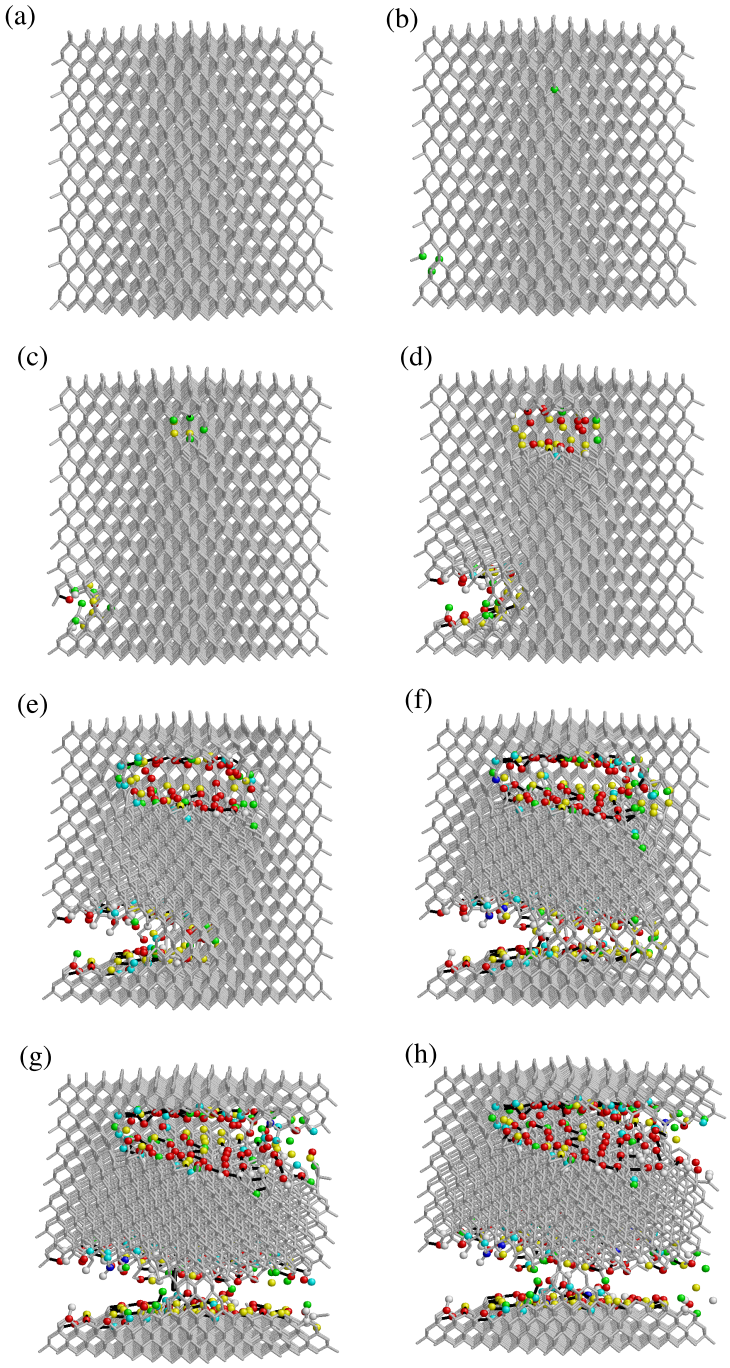


Figure 7.26: Snapshots of the dynamical fracture simulation of the '4501B' sample with 4501 Si atoms. The present figures (a) ,(b)... (h) show, in the view method of Fig. 7.24(β), the same snapshots in Fig. 7.25 (a) ,(b)... (h), respectively,

Anisotropic fracture propagation mechanism

Now we discuss the anisotropic fracture propagation mechanism on a flat (001) surface. The anisotropy is seen in the '4501' sample, especially within the early snapshots of the Fig. 7.21. The bond-breaking propagation is anisotropic among the [110] and [1$\bar{1}$0] directions.

This anisotropy is due to the quantitative difference of the successive bond breaking mechanism between the two directions, as explained below. A schematic picture is shown in Fig. 7.27. In the [110] direction, the successive bond breakings propagate along the *nearest neighbor* bond sites, which forms a connected zigzag path, as in Fig. 7.27(a). Since the stability of the sp^3 bonding state is strictly limited to the four-fold coordination, a bond breaking process significantly weakens the *nearest neighbor* bond sites. Therefore, the bond-breakings tend to occur at successive bond sites simultaneously and the resultant two-fold coordinated atoms are stabilized by the dehybridization mechanism, as explained in Fig. 7.6(a)-(c). This mechanism means a local electronic instability and we call this fracture mode 'electronic' mode. In the [1$\bar{1}$0] direction, on the other hand, the bond-breaking paths are not connected as in Fig. 7.27(b). In this direction, the bond breakings are propagated through the local strain relaxation, as in the Griffith theory (See Section 6.1). We call this mode 'elastic' mode.

In the 'elastic' mode, the strain relaxation mechanism requires the atomic motion, while, in the 'electronic' mode, the electronic instability can propagate without atomic motions. As results, the bond breaking propagation along the *nearest neighbor* bond sites (in the [110] direction of the present surface) can be faster than that in the perpendicular direction (in the [1$\bar{1}$0] direction). Moreover, in Section 7.3, we discussed that the formation of a cleaved surface does not occur without the dehybridization mechanism, which is equivalent to the crucial role of the 'electronic' mode at the early stages of the present fracture phenomena.

Comment on other features

Now we make comments on other features seen in the results of the '4501A' and '4501B' samples; (i) In the '4501A' sample, another fracture region seems to be created from the crack tip, which is seen among the snapshots of Fig. 7.22(d)-(h). In the continuum theory of fracture, the crack tip shows the singularity in the stress field (See Appendix C.2). Several related topics, such as dislocation emissions, are focused in the atomistic theory. The above feature might be interesting in the above context. Moreover, one may think that the cleaved plane is bended into the (111) plane at the left-down corner of the above snapshots. We should say, however, that these features are seen on the region near a sample edge. Further investigations should be done with different sample geometries or sizes, which are possible future works. (ii) In the '4501B' sample, a domain with several dimer rows seems to be formed, which is seen in Fig. 7.24. From the classification in Section 6.3, the flipping freedoms of the asymmetric dimers in the domain are in the (2×1) configuration, which shows the anisotropy between the [110] and [$\bar{1}\bar{1}$0] directions. This anisotropy is consistent to the anisotropy due to the fracture propagation direction; the fracture propagates in the [110] direction and the [110] and [$\bar{1}\bar{1}$0] directions are inequivalent due to the presence or the absence of the crack tip. We discussed in Section 6.3

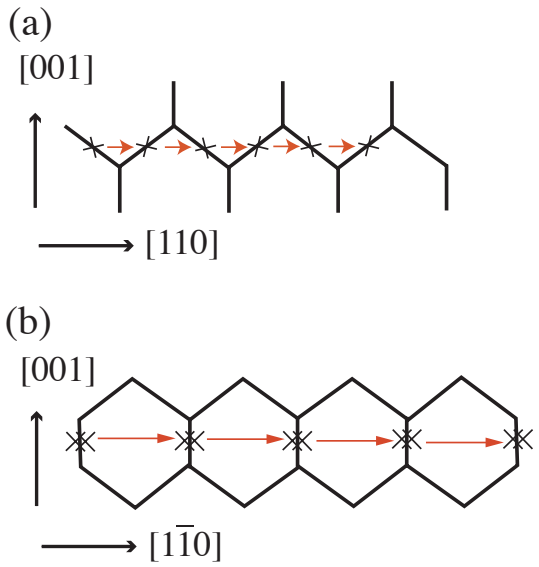


Figure 7.27: The anisotropic fracture propagation mechanism within a (001) plane; (a) 'electronic' fracture mode, in which the fracture path forms a connected (zigzag) line. (b) 'elastic' fracture mode, in which the fracture path does not forms a connected line. In (a), a broken bond site is labeled by the *single* cross, which indicates a single bond site. In (b), a broken bond site is labeled by the *double* cross, which indicates successive zigzag bond sites that lies perpendicular to the paper.

that the local environment may affect the flipping motions. This feature may be interesting in this context and further investigations should be done as one of possible future works. We should be careful, however, to discuss the flipping motions of the asymmetric dimers, because such motions can be activated in fine energy scales, as shown Table 6.1.

7.6 Fracture simulation with 10^4-10^5 atoms

In this section, the fracture simulations are done with $10^4 - 10^5$ atoms, which are larger samples than those in the previous section (Section 7.5). As a characteristic feature of the present large samples, several step structures will be discussed.

Two samples are prepared, which contains the initial defect bond as the fracture seed. One sample has the size labeled by $n_{110} \times n_{1\bar{1}0} \times n_{001} = 49 \times 50 \times 49$, in the unit of the number of atomic layers. Here $n_{110} = 50$ corresponds to about 10 nm. The sample contains 30025 atoms and is denoted as '30025A' sample. The other sample has the size labeled by $n_{110} \times n_{1\bar{1}0} \times n_{001} = 97 \times 100 \times 49$. The sample contains 118850 atoms and is denoted as '118850A' sample. Within the atoms on the 'top' or 'bottom' sample surfaces, the z component of their motion is under the constraint of the constant-velocity motion, and the x and y components are frozen. Within the atoms on the 'side' sample surfaces, the x and y components of their motion are frozen. Note that the above condition is just one of reasonable ones. The effect of different conditions on the fracture behavior might be one of the possible future works. The main purpose of the present simulations is the sample size dependence of the fracture behavior. Therefore, the important point is the fact that, in the two cases, all the conditions are the same, except the sample size.

We use the hybrid scheme between the variational and perturbative order-N methods. The variational method is used only for selected Wannier states whose localization centers are located near the fracture region. Some of such wave functions change their character dynamically from the bulk (sp^3 bonding) states to surface ones, as discussed in Section 7.2. We choose the Wannier states for the variational method, if the z coordinate of its localization center locates within ± 4 atomic (bond) layers from that of the initial defect bond site. In results, the Wannier states whose localization centers lie among the nine bond layers are treated in the variational method. The above region is sufficient for the fracture simulation, because the fracture will occur only within ± 2 atomic (bond) layers from that of the initial defect bond site, as discussed below. In our program code, the member of the selected Wannier states can be dynamically controlled, as explained in Section 7.4. In the present simulations, however, the above member of the selected Wannier states is unchanged during the simulation. In the '118850A' sample, the following additional condition is applied to the selection of the Wannier states; the localization center of the Wannier state must be placed, on the xy plane, within a circular region whose center is the initial defect bond site. The circular region will be shown in the figures of this section. As results, about 10^4 Wannier states are treated in the variational order-N method.

Results of the sample with 30025 atoms

Figures 7.28, 7.29, show the result of the '30025A' sample. Figure 7.28 is plotted in the same manner as in the previous section; Each Wannier state is classified from its weight distribution into a bonding or atomic orbital, which is shown as a rod or a ball in the figures, respectively, To see the step structures clearly, Fig. 7.29 is plotted in a different manner for the same snapshot; The broken bond sites are shown as colored rods in the *ideal* crystalline geometry. The color of rods indicates

the atomic layer as shown in the figure with samples. Especially, the atomic layer of the *black* rods is the layer that contains the initial defect bond. Figures 7.30, 7.31, 7.32 show the successive snapshots of the dynamical fracture simulation.

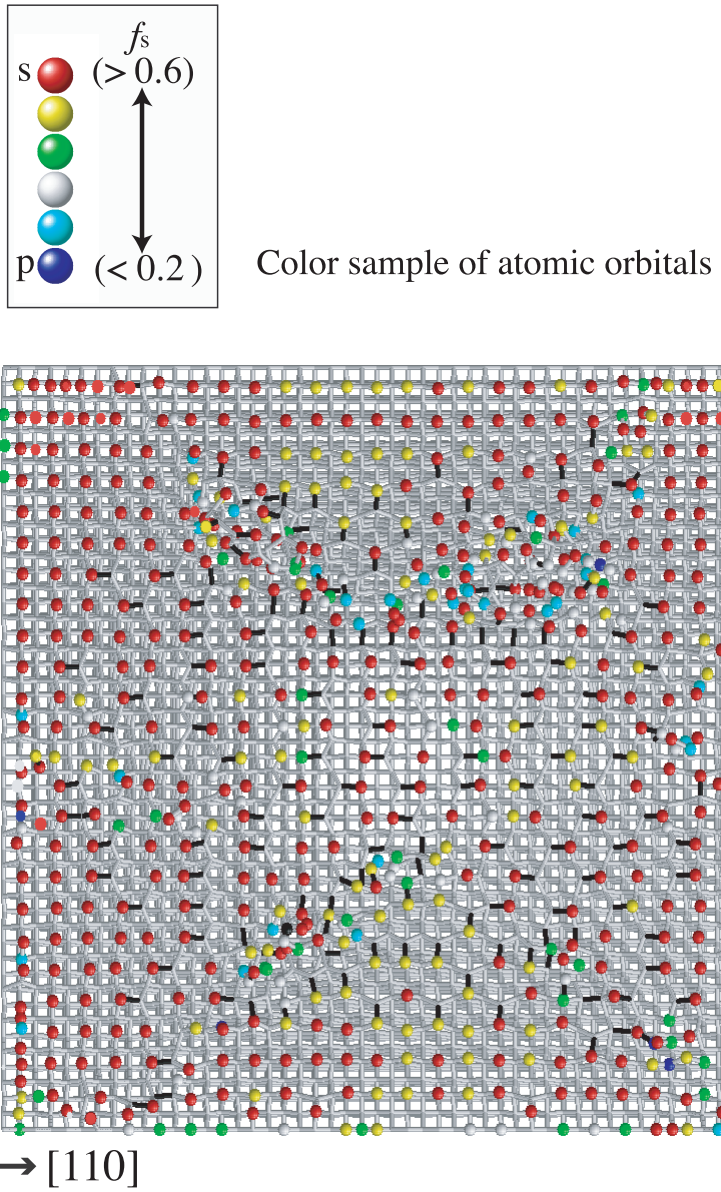


Figure 7.28: Top view of the cleaved plane in the fractured '30025A' sample with 30025 Si atoms. See the caption of Fig. 7.19 for the descriptions of the rods and the balls.

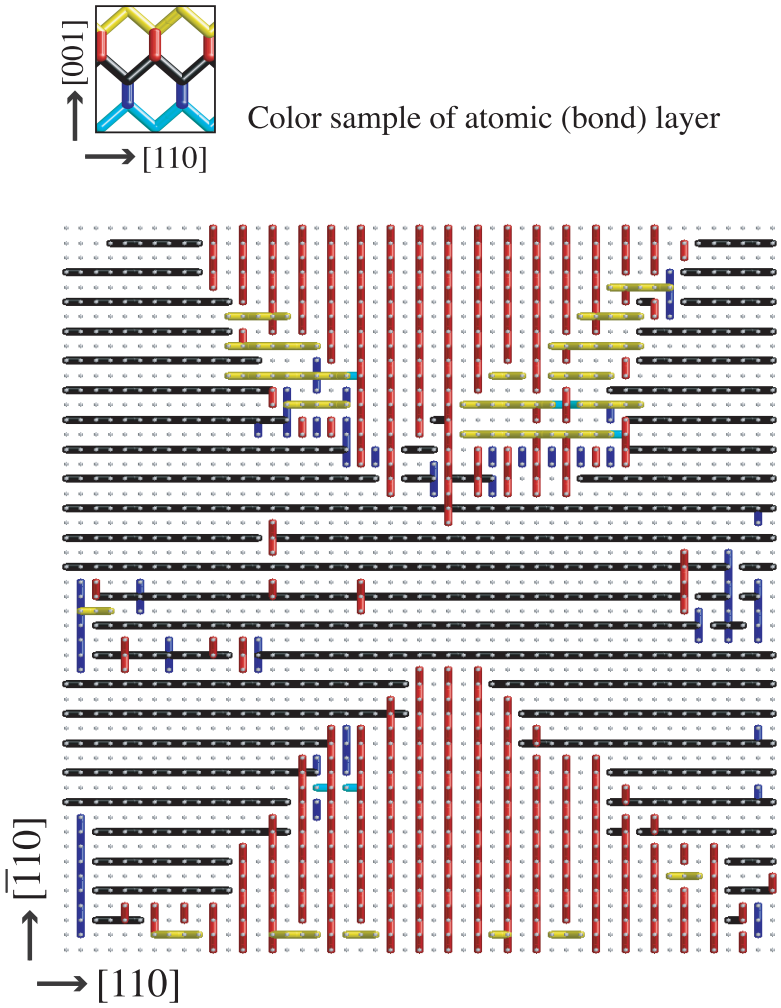


Figure 7.29: Fracture geometry of the '30025A' sample with 30025 Si atoms. The broken bond sites are plotted as colored rods in the ideal (crystalline) geometry. The color of rods indicates the atomic layer. The color sample is shown in the small figure.

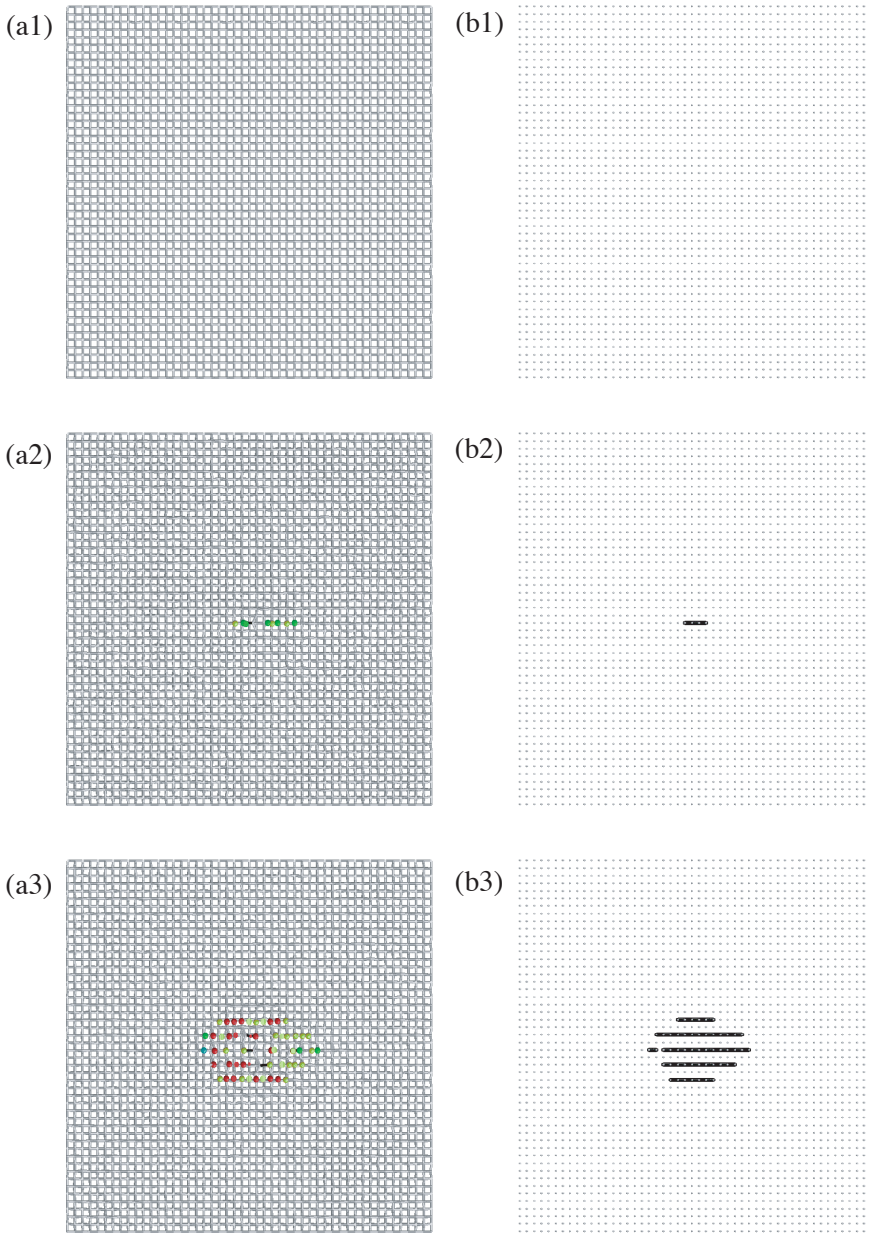


Figure 7.30: The snapshots of the dynamical fracture simulations of the '30025A' sample. The figures labeled '(a)' are plotted in the manner of Fig. 7.28, while the figures labeled '(b)' are plotted in the manner of Fig. 7.29. The number in the figures indicate the time. For example, the figures (a1) and (b1) shows the same snapshot in different plotting manner. The time interval between the two successive snapshots is $\Delta t = 0.225$ ps.

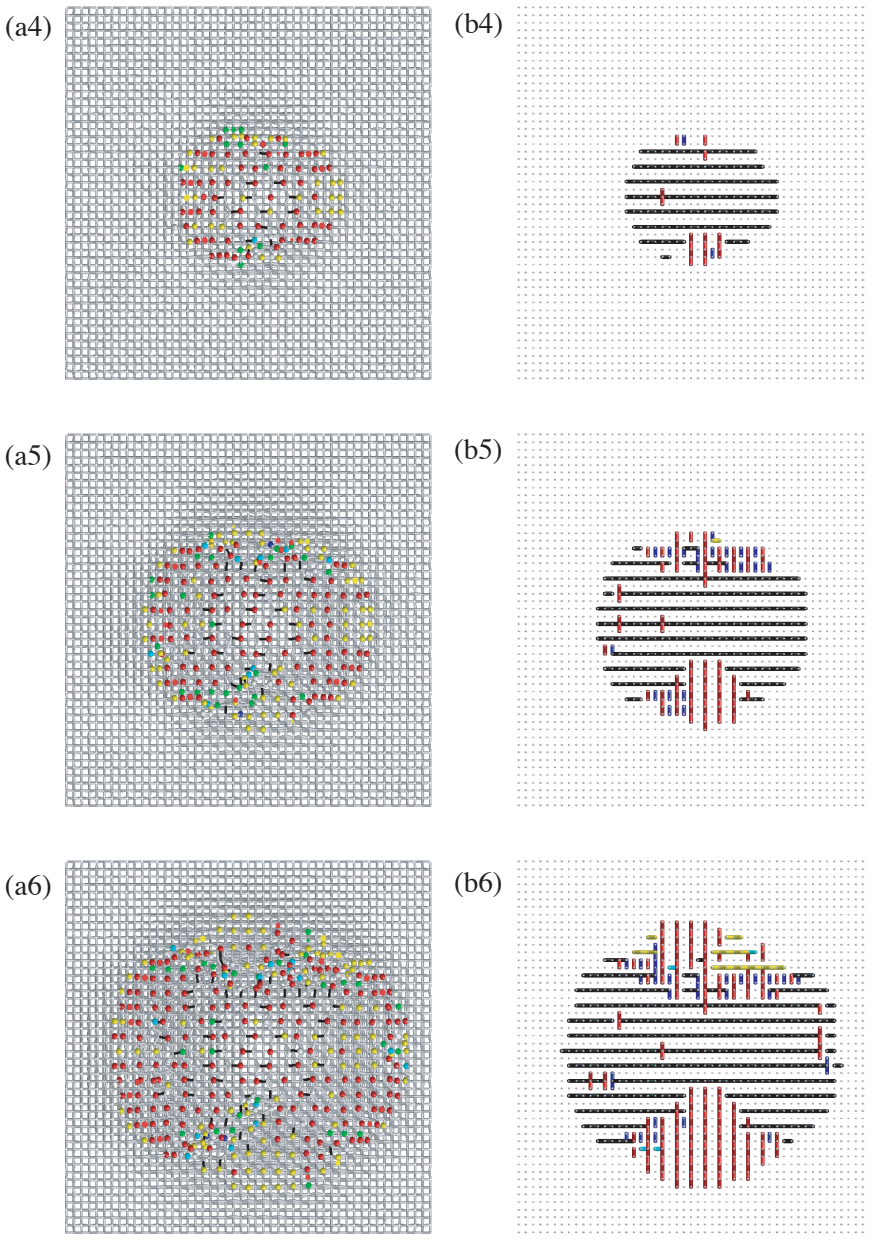


Figure 7.31: The snapshots of the dynamical fracture simulations of the '30025A' sample. These snapshots are continued from Fig.7.30. See Fig.7.30 for details of the plotting manner.

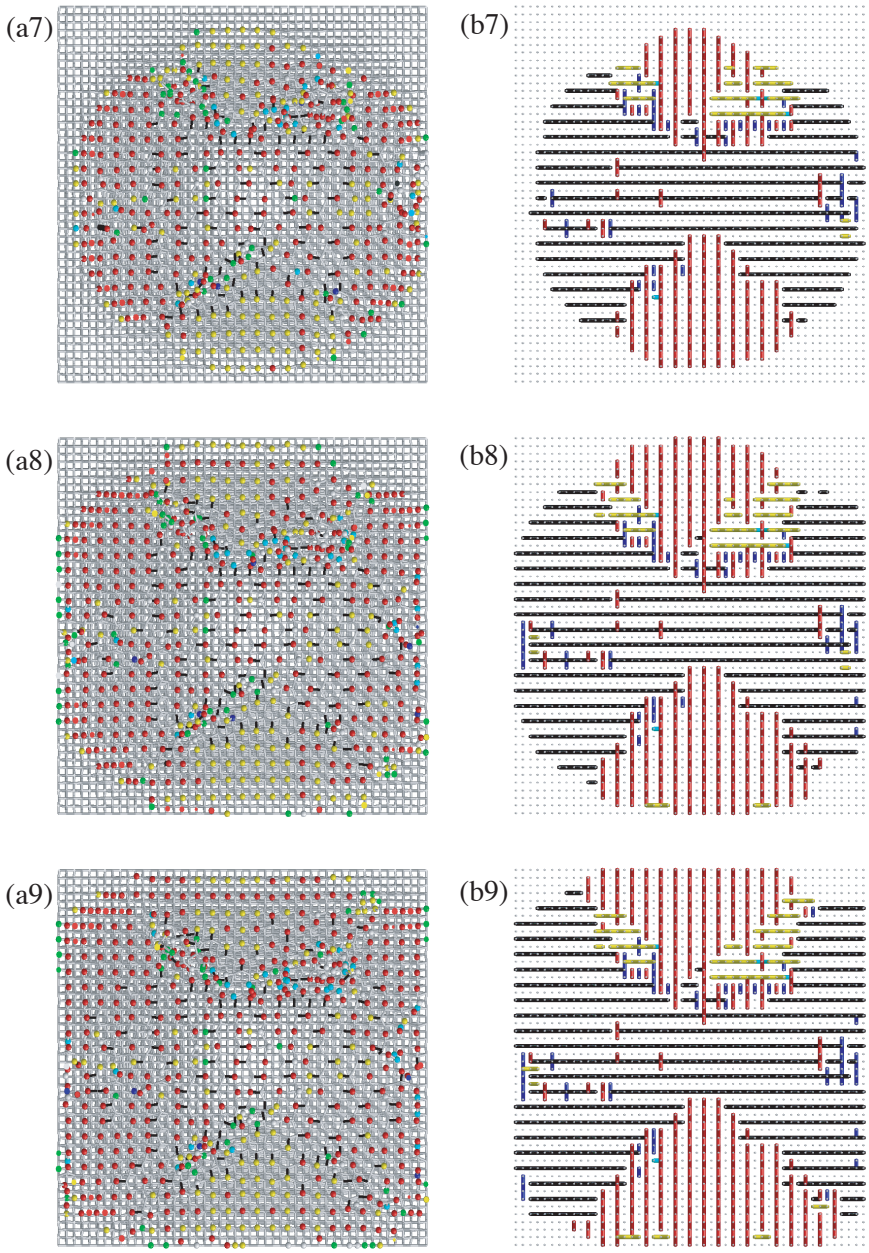


Figure 7.32: The snapshots of the dynamical fracture simulations of the '30025A' sample. These snapshots are continued from Fig.7.31. See Fig.7.30 for details of the plotting manner. Fig. 7.28 and Fig. 7.29 correspond to the snapshots (a10) and (b10), respectively.

Results of the sample with 118850 atoms

Similar figures are also shown in the '118850A' sample, in Figs. 7.33, 7.34, 7.35, 7.36, 7.37. Unlike the figures in the '30025A' sample, the figures in the '118850A' sample shows only a central area ($n_{110} \times n_{1\bar{1}0} = 58 \times 60$). Moreover, the visualization of Wannier states as rods or balls is done only for the selected Wannier states in the variational method. From the initial preparation, the area of visible rods or balls forms a circular area whose center is the initial defect bond site. Within the resultant figures, we can see that the fracture propagate almost in a circular region whose center is the initial defect bond site.

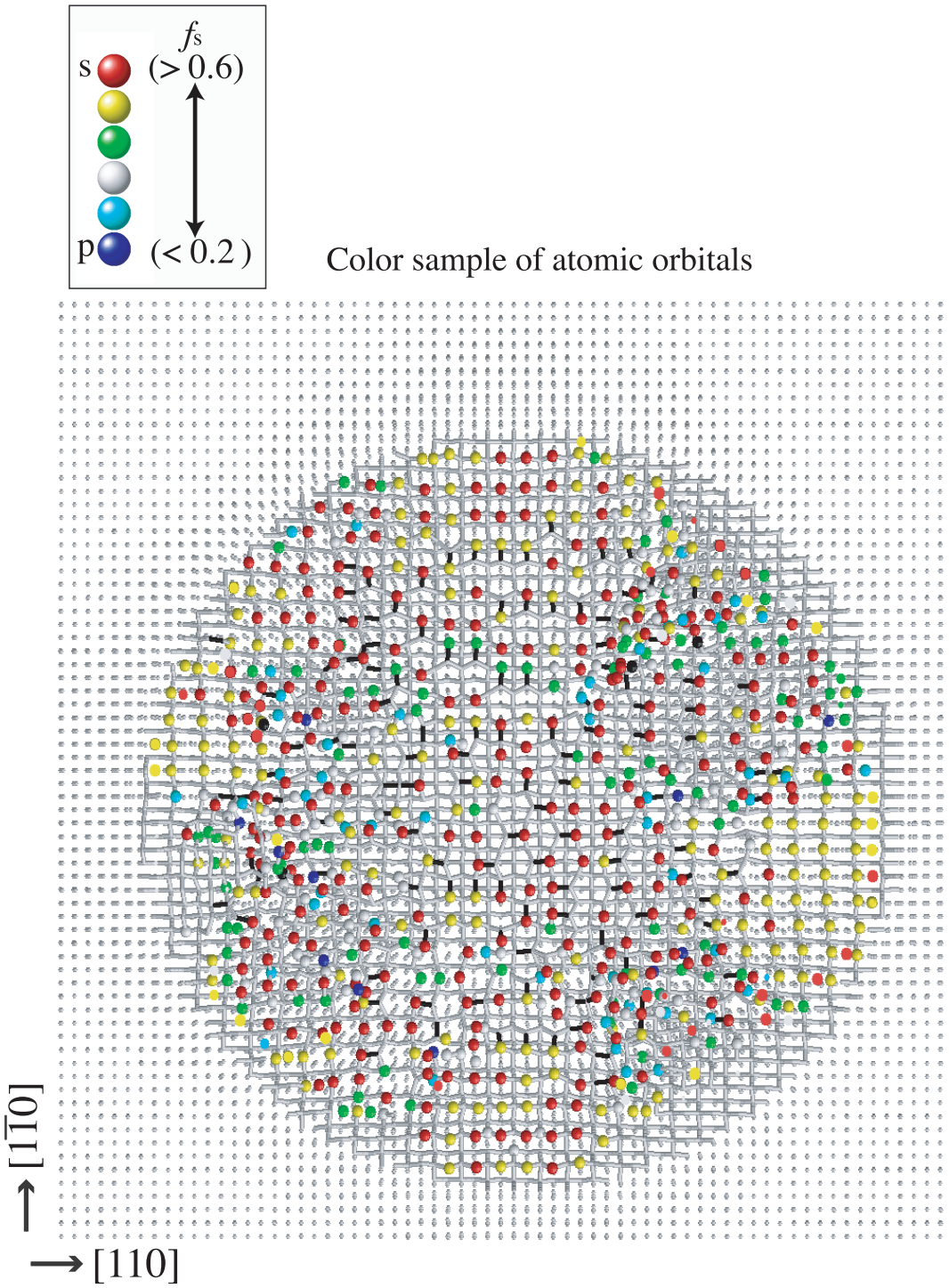


Figure 7.33: Top view of the cleaved plane in the fractured '118850A' sample with 118850 Si atoms. The figure shows only a central area ($n_{110} \times n_{1\bar{1}0} = 58 \times 60$), while the whole sample is labeled by $n_{110} \times n_{1\bar{1}0} \times n_{001} = 97 \times 100 \times 49$. See the caption of Fig. 7.19 for the description of the rods and the balls. Here the visualization of rods or balls is limited in a circular area whose center is the initial defect bond site.

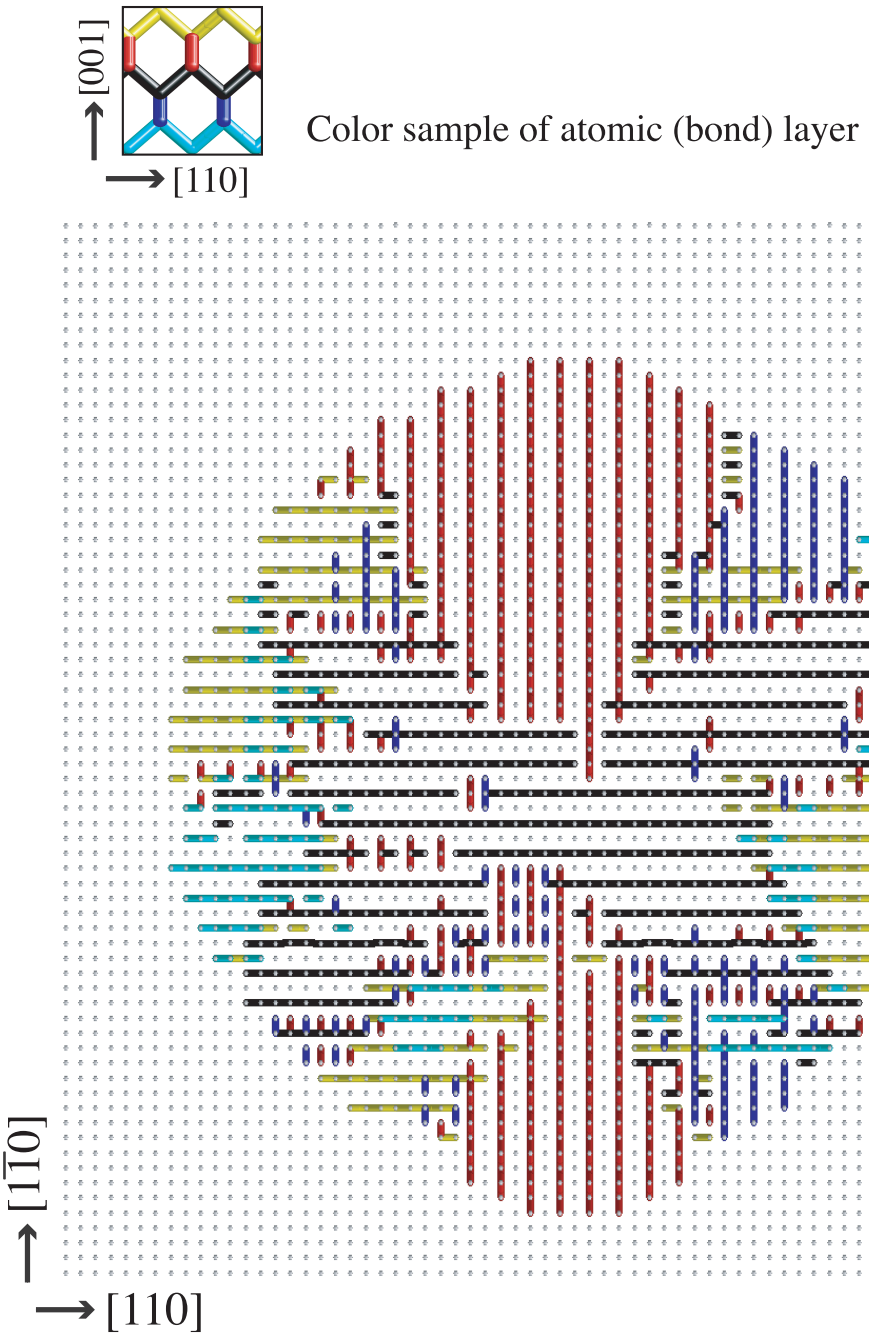


Figure 7.34: Fracture geometry of the '118850A' sample with 118850 Si atoms. The figure shows only a central area ($n_{110} \times n_{1\bar{1}0} = 58 \times 60$), while the whole sample is labeled by $n_{110} \times n_{1\bar{1}0} \times n_{001} = 97 \times 100 \times 49$. See the caption of Fig. 7.29 for the plotting manner. Note that the length of $n_{110} = 50$ atomic layers is about 10 nm.

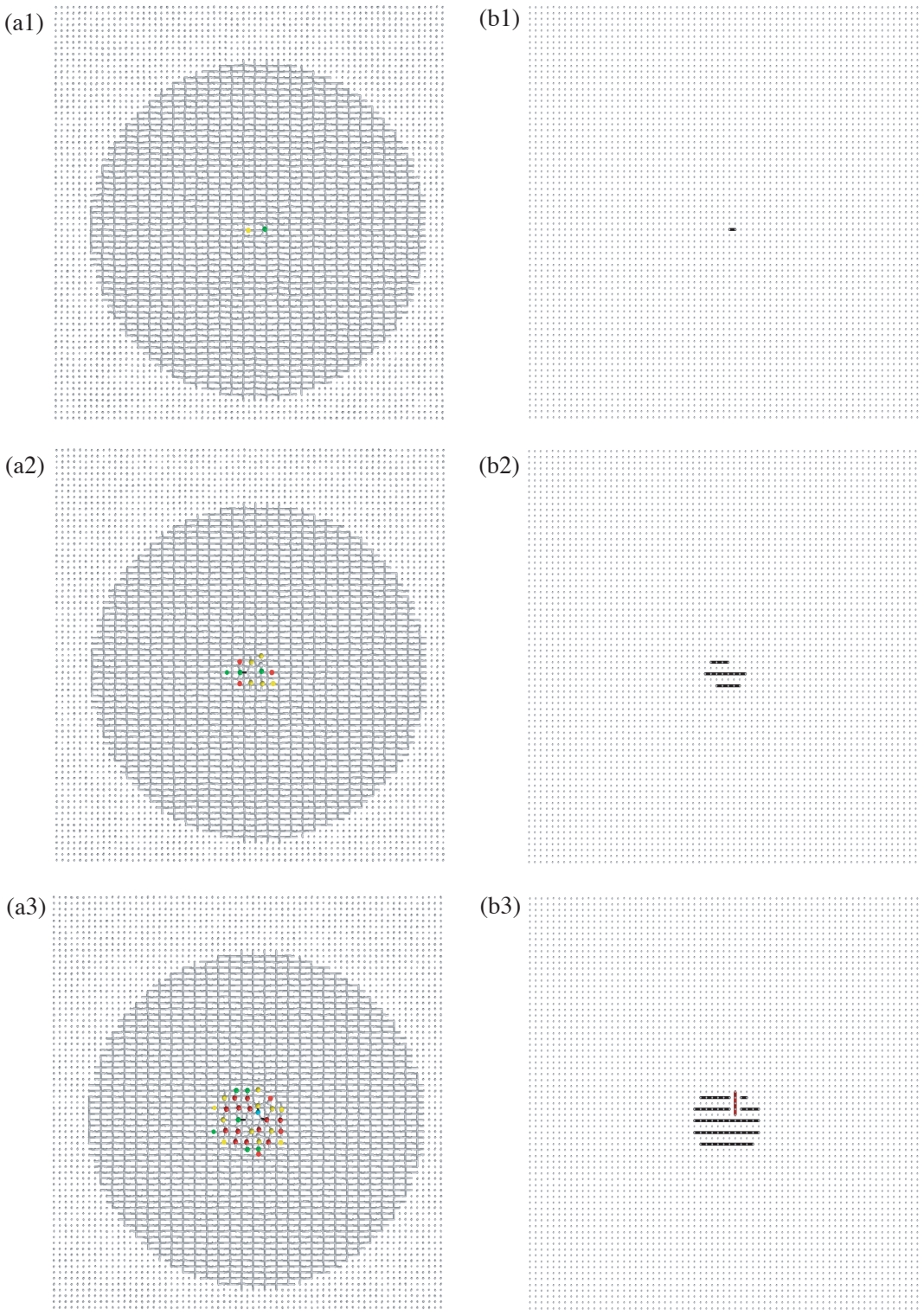


Figure 7.35: The snapshots of the dynamical fracture simulations of the '118850A' sample. The figure shows only a central area ($n_{110} \times n_{1\bar{1}0} = 58 \times 60$), while the whole sample is labeled by $n_{110} \times n_{1\bar{1}0} \times n_{001} = 97 \times 100 \times 49$. See Fig.7.30 for the plotting manner. The time interval between two successive snapshots is $\Delta t = 0.15$ ps.

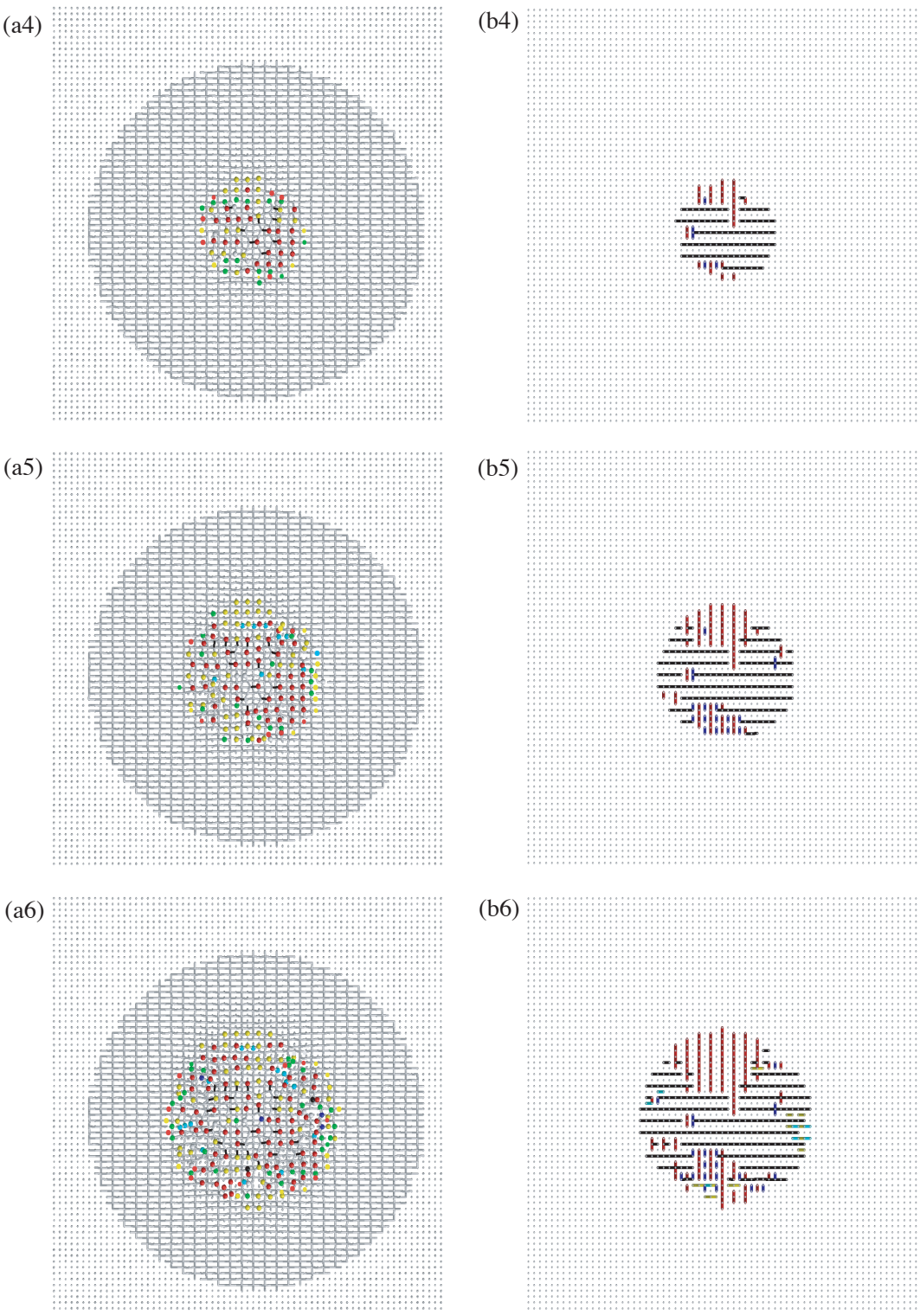


Figure 7.36: The snapshots of the dynamical fracture simulations of the '1188505A' sample. These snapshots are continued from Fig.7.35. See Fig.7.35 for the plotting manner.

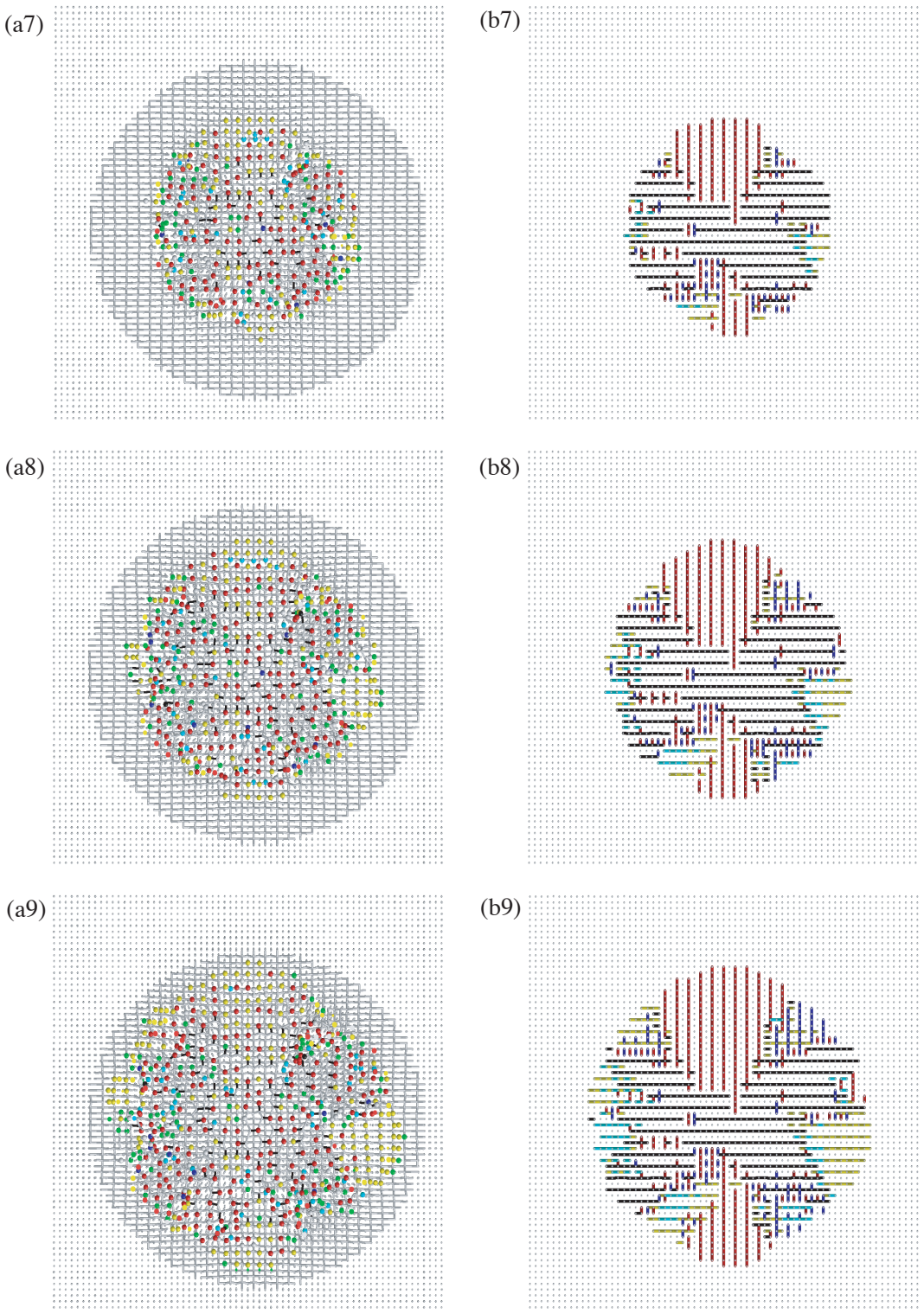


Figure 7.37: The snapshots of the dynamical fracture simulations of the '1188505A' sample. These snapshots are continued from Fig.7.36. See Fig.7.35 for the plotting manner. Fig. 7.33 and Fig. 7.34 correspond to the snapshots (a10) and (b10), respectively.

Step structure of cleaved surface

Here the mechanism of the step formation is discussed. The elastic property of silicon crystal shows only a small anisotropy within (001) plane; the [110] and [$\bar{1}$10] directions are equivalent and the values of the Young modulus are different by only about 30 % in the [100] and [110] directions (See Appendix B.1). On the other hand, as explained in Section 7.5, the anisotropic bond-breaking propagation in one (001) plane increases the anisotropic strain energy. The anisotropy originates from the inequivalence between the [110] and [1$\bar{1}$0] directions within *one* (001) layer. Since the above inequivalence does not appear within *two* successive layers, a step formation between them will release the anisotropic strain energy. In the '30025a' sample, a step is formed between the layer of black rods and that of red rods. In the [110] direction, the bond-breaking propagation reaches the sample surfaces without step formations. In the [1$\bar{1}$0] directions, the bond-breakings propagate slower within the early snapshots and a step is formed in the central area. After that, the fracture propagates among the two atomic layers. Since the anisotropic fracture propagation mechanism is symmetrically equivalent among the two layers, the resultant step formation path almost forms lines in the [100] and [010] directions, as the boundary between the fractured areas within the two layers.

In Fig.7.34, the largest sample in the present thesis, the above line structure does not reach the sample surface but is canceled with additional step formations in complicated paths. The sample size dependence of the step structures is understood by the beginning of the crossover between nanoscale and macroscale samples; If the sample contains so many atoms, the geometry of the resultant crack will be almost circular, as in Fig. 7.34, so as to minimize the anisotropic strain energy. If not, the strain energy is accumulated only within the confined bulk region due to the finite sample size. The resultant fracture behavior is directly related to the anisotropic atomic structure of the cleaved surface, as in Fig. 7.29.

Since the above mechanism of step formations is two dimensional, the present samples may be nanoscale 'thin' samples. The 2D-like situation can be seen in Fig. 7.38, in which the deformations are frozen at the 'top' and 'bottom' sample surfaces, due to the boundary condition. In larger or thicker samples, an expected fracture behavior is the bending of the fracture plane into the (111) plane, the easiest cleavage plane in macroscale samples, which is the crossover in the present context. Note that the deviation of the cleaved area from the (001) plane was seen in Fig. 7.20, though it is near the sample edge. In an *sufficiently* large sample, the fracture mode with the easiest cleavage plane will grow regardless of sample shape and details of conditions.

Note that the dynamical simulation with 10^5 atoms is a practical limitation within a single CPU workstation, since the total simulation time is more than one week. We will discuss the future aspect in the next section.

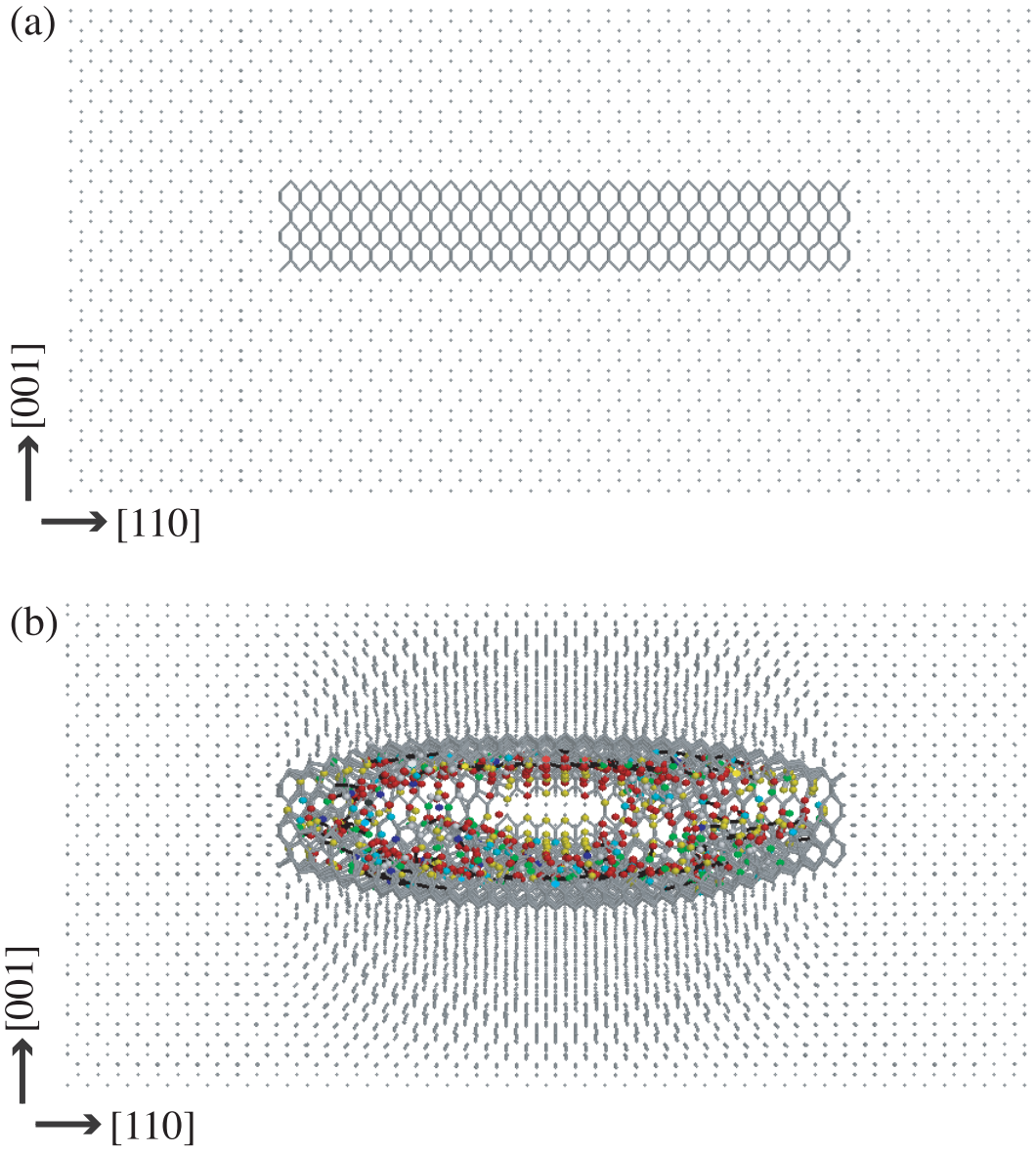


Figure 7.38: A 'semi-infinite' view of the '118850A' sample. See Fig.7.33 for the plotting manner. The figure (a) and (b) correspond to the initial and final (fractured) snapshots.

7.7 Summary and discussion

Summary

Fracture processes of the nanocrystalline silicon were simulated among 10^2-10^5 atoms under the [001] external load. We focused on the two issues as the purposes of the simulations (Section 7.1). The two issues were investigated with the analysis of quantum mechanical freedoms of electron systems. The following points were discussed;

(I) When the fracture begins, the electronic energy gap vanishes as level crossings between several 'defect' states. Here 'defect' states should be interpreted as electronic states that do not appear in bulk (crystalline) structure. In fractured samples, these 'defect' levels form a surface band (Section 7.2).

(II) The fracture process is observed with Wannier states. The elementary fracture process accompanies the two-stage surface reconstruction. First, the two-fold coordinated atom with a lone pair state is stabilized by the dehybridization mechanism. Then, a pair of two-fold coordinated atoms form an asymmetric dimer (Section 7.2). Without the dehybridization mechanism, no cleaved surface is formed. (Section 7.3).

(III) Due to the anisotropy within a (001) single atomic layer, there are two kinds of the fracture propagation mechanisms. We call them the 'electronic' mode and the 'elastic' mode. The former mode is directly related to the dehybridization mechanism (Section 7.5).

(VI) Steps are formed within larger samples. The origin of step formations can be explained by releasing the anisotropic strain energy within the (001) plane, which corresponds to the beginning of the expected crossover between the nanoscale and macroscale samples (Section 7.6).

These results show the crucial importance of the quantum mechanical freedoms of electron systems, particularly the dehybridization mechanism. These results also show the crucial importance of large-scale calculations, because the fracture behaviors are essentially different among the sample sizes.

Discussion

Hereafter, we discuss several possible future works within the fracture of silicon (nano)crystals.

For simulations with larger samples and/or longer timescales, the program code in parallel computations should be prepared, as discussed in the last paragraph of Chapter 7.6. We have finished the parallelization of the perturbative order-N method (Section 5.4). For the parallel computations of fracture simulations, the parallelization of the variational order-N method is also required. As in the perturbative order-N method (Section 5.4), the dominant procedures in the variational order-N method can be parallelized with respect to the Wannier states, which is seen in the chart of Fig. 5.1 (Section 5.2). One important technical point is the choice of the parallelization technique. We have tested the two standard techniques, the MPI

technique and the OpenMP technique, for the perturbative order-N method (Section 5.4). We will determine which is suitable for the variational order-N method.

When the simulation method with the parallel computations is established, several systematic investigations will be done among different conditions, such as different sample sizes, sample shapes, and boundary conditions.

Even with parallel computations, however, it is *impractical* to calculate the macroscale fracture phenomena, with 10^{23} atoms, using the present large-scale electronic structure methods, Therefore, the theory should be mapped to the continuum theory in a reasonable scheme. Since the present method is based on the well-defined total energy functional, the mapping will be done as the mapping of the total energy functional.

Chapter 8

Summary and general discussion

Summary

In this thesis, we constructed the theory for large-scale simulations by simplifying the total energy functional in the electronic structure theory. Its foundation is based on (a) the tight-binding Hamiltonian, especially its universality, (b) several order-N methods, mainly those with the generalized Wannier states and (c) the hybrid scheme by dividing the Hilbert space. The summary and general discussions for the theories were done in Section 3.3 and Section 4.5. Test calculations are done up to 10^6 atoms with or without parallel computers. As a practical large-scale calculation, the fracture process of nanocrystalline silicon was simulated with up to 10^5 atoms. The results show that the quantum mechanical freedoms of the electron systems is crucial for the fracture mechanism, as summarized in Section 7.7.

It is important that the above three theories are well defined within the quantum mechanics. Therefore, the applicability is *not* limited to silicon or fracture simulation. It is also important that the above three theories are independent, not only as the theoretical concepts, but also as the corresponding subroutines in the program code. We can develop these theories or subroutines independently. Since the present program code consists of more than ten thousand lines, the independence among subroutines is important, especially, when several persons join the development of the program code.

By combining the three theories, we will design various simulations, among different numbers of atoms, different computational costs, and/or different systems.

General discussion with 'multiscale mechanics'

Apart from the methodology, now we discuss the possible targets of large-scale electronic structure calculations. In Chapter 1, we discussed the concept 'multiscale mechanics', as a simulation method with different *length* scales. Here we discuss several phenomena with different *energy* scales.

As an overview of the present work, Fig. 8.1 shows the typical energy scales in silicon. The interactions in relatively low energy scales are those of the flipping freedoms of asymmetric dimers on a clean (001) surface, which was explained in Section 6.3. Some of their energy scales are smaller than the kinetic energy at room temperature, which can be seen, for example, by low temperature STM observations. In a general theoretical viewpoints, phenomena with a fine energy scale are difficult

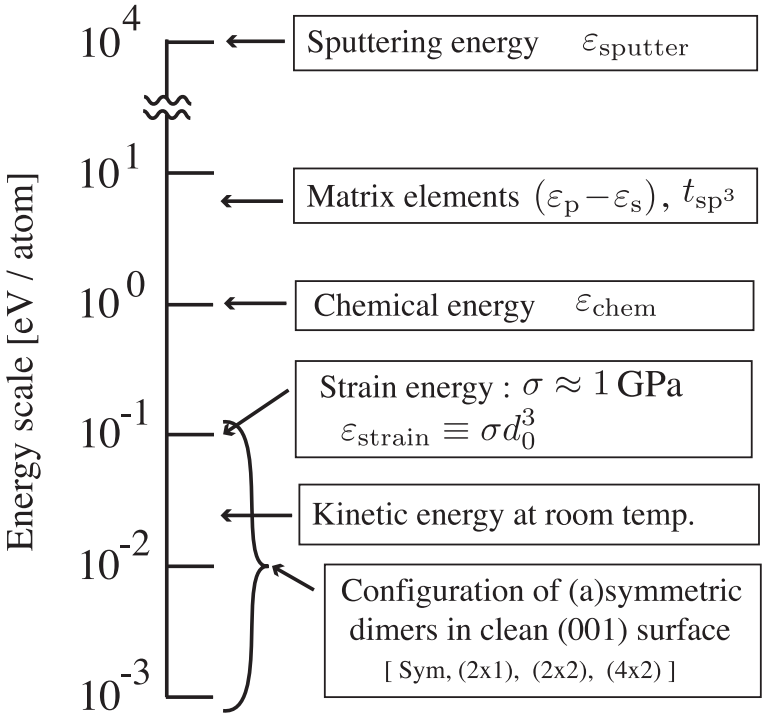


Figure 8.1: Various energy scales in silicon.

to reproduce by the total energy method of electronic structures. The interactions in an 'intermediate' energy scale, energy scales in 1-10 eV, are directly related to the universal tight-binding theory in Chapter 3. These energy scales characterize the diagonal and off-diagonal elements of the Hamiltonian matrix. As the competition of the above energies, the bonding and dehybridization mechanisms are characterized by the energy scale of 1 eV ($\varepsilon_{\rm chem}$). As a relatively high energy scale, the sputtering energy can be considered, typically, in the scale of $\varepsilon_{\rm sputter} = 10$ KeV.

Realistic materials contain various interactions among various energy scales. An important point for constructing a proper theoretical model is whether the decoupling treatment between interactions is justified or not. In many situations, two interactions in different energy scales are decoupled. An example is the frozen core approximation in electronic structure calculations. This is justified, when the binding energy of core electrons is much larger than the cohesive energy of valence electrons. Several continuum or classical models are derived with decoupling treatments. In perfect Si crystal at room temperature, for example, bond breaking processes are not expected and only elastic motions are expected. This is the decoupling treatment based on the fact that the bondbreaking energy ($\varepsilon_{\rm chem}$) and the thermal energy ($\varepsilon_{\rm thermal}$) are quite different in the energy scale ($\varepsilon_{\rm chem} \gg \varepsilon_{\rm thermal}$). This decoupling treatment justifies a continuum or classical model that reproduces the linear elasticity, but not bond breaking processes.

In several situations, however, two interactions in different energy scales are *not* decoupled. Suppose a situation in which two energy terms are competitive in the total energy, though they are quite different in *length and energy* scales. Such situation may occur in inhomogeneous systems. One example is the sputtering phenomena,

in which one atom with a huge kinetic energy, typically $\varepsilon_{\rm sputter} = 10$ KeV, possibly causes the bond breakings, characterized by $\varepsilon_{\rm chem} = 1$ eV, among many atoms. In other words, a *small* number of atoms has a *large* energy ($\varepsilon_{\rm sputter}$), while a *large* number of atoms has a *small* energy ($\varepsilon_{\rm chem}$). The two energy terms can be competitive in the total energy, which does *not* justify the decoupling treatment. Here we say that such a system has the 'multiscale feature', in the sense that the system essentially contains two competitive interactions in different energy and length scales. For such systems, a critical size, or a critical number of atoms, n should be defined as the ratio between the energy scales of the interactions. In the sputtering phenomena, the critical number of atoms is given as

$$n = \frac{\varepsilon_{\rm sputter}}{\varepsilon_{\rm chem}} = 10^4. \tag{8.1}$$

The above 'multiscale feature' originates from the inhomogeneous property of the kinetic energy density, which will not appear in a thermal equilibrium system, due to the equipartition theorem. The brittle fracture is another example of the above 'multiscale features', due to the competition between the chemical energy and the strain energy. The ratio

$$n = \frac{\varepsilon_{\rm chem}}{\varepsilon_{\rm strain}} \tag{8.2}$$

corresponds to the Griffith critical crack length for fracture. In this case, the system has an inhomogeneous structure, due to the presence of cleavage planes in a bulk sample. As discussed in Section 6.1 or Section 6.3, the theory of nucleation or the step formation mechanism of Si(001) surfaces can be explained by the analogous energy competition mechanism in inhomogeneous structures.

In general, the critical size for the energy competition n is independent on the system size N. Therefore, one can expect a crossover among the system sizes N, in which the critical system size is given as $N \approx n$. In this thesis, we discuss such crossovers among the nucleation mechanism, the fracture mechanism and the step formation mechanism (See Section 6.1 and Section 6.3). Such a crossover can be a target of large-scale electronic structure calculations, because quantum mechanical processes of electron systems are essential and the system size of the simulation should be large enough for the critical size n.

The *time scale* is also important to characterize physical phenomena. In the fracture simulations of the present thesis, the total simulation time is on the order of pico second. The time scale is governed by the crack propagation speed, which is on the order of the sound or Rayleigh wave speed. Several phenomena, such as growth, are in a *much longer* time scale than those discussed above. To simulate such *really long-time* phenomena, several fundamental theories are desirable but are not included in the present thesis. These methods should be a future work in the general context of the process simulation.

We classified physical phenomena by their typical length, energy and time scales. Such classification may give a guiding principle for further developments and applications of large-scale electronic structure calculations.

Appendices

Appendix A

Note on electronic structure theory

A.1 First-principle molecular dynamics and limitation of LDA

Based on the explanation in Section 2.1, this appendix describes two topics in *ab initio* theory; the first-principle molecular dynamics and the limitation of the LDA.

First-principle molecular dynamics

The first-principle molecular dynamics [3] is based on the DFT energy functional, Eq. (2.2), under the adiabatic (Born-Oppenheimer) approximation. The method is also based on the *ab initio* pseudo potential theory [14, 15, 16, 17], in which the wave functions only for the *valence* electrons are explicitly treated as 'pseudized' ones. The guiding principle for generating pseudo potentials is to reproduce the scattering property within the linear order of energy. 'True' valence wave functions oscillate in the core region, the region near the atomic nucleus, due to the orthogonality to the core wave functions. The 'pseudo' wave functions, on the other hand, does not have the oscillating behavior at the core region. Such 'smooth' wave functions can be expanded by a relatively small number of plane wave bases.

In the first-principle molecular dynamics, a periodic simulation cell is used and each (pseudized) wave function is expanded by plane wave bases as

$$\phi_i(\boldsymbol{r}) = \sum_{\boldsymbol{g}}^{|\boldsymbol{g}|<g_{\rm c}} c_{i\boldsymbol{g}} e^{i\boldsymbol{g}\boldsymbol{r}}. \tag{A.1}$$

The reciprocal vectors ($\{\boldsymbol{g}\}$) are defined for the simulation cell. Here, for simplicity, the wave functions ϕ_i are limited to those at the Γ point ($\boldsymbol{k} = 0$) in the Brillouin zone. The plane wave bases $\{e^{i\boldsymbol{g}\boldsymbol{r}}\}$ in Eq.(A.1) are limited within a cutoff wave number $g_{\rm c}$ ($|\boldsymbol{g}| < g_{\rm c}$). The corresponding kinetic energy ($g_c^2/2$) is usually called 'cutoff energy'. The wave functions are stored as the set of the plane wave coefficients $\{c_{i\boldsymbol{g}}\}$. The kinetic energy term is written as

$$E_{\rm kin} = \sum_i^{\rm occ.} \langle\phi_i| \frac{-\nabla^2}{2} |\phi_i\rangle = \sum_i^{\rm occ.} \sum_{\boldsymbol{g}}^{|\boldsymbol{g}|<g_{\rm c}} \frac{1}{2} g^2 |c_{i\boldsymbol{g}}|^2. \tag{A.2}$$

Using Eq. (A.2) and a mathematical relation

$$\begin{aligned} \frac{\partial}{\partial c_{i\boldsymbol{g}}^*} &= \int d\boldsymbol{r} \frac{\partial \phi_i^*(\boldsymbol{r})}{\partial c_{i\boldsymbol{g}}^*} \frac{\delta}{\delta \phi_i^*(\boldsymbol{r})} \\ &= \int d\boldsymbol{r} e^{-i\boldsymbol{g}\boldsymbol{r}} \frac{\delta}{\delta \phi_i^*(\boldsymbol{r})}, \end{aligned} \tag{A.3}$$

the energy gradient with respect to one coefficient is given by

$$\frac{\partial E_{\rm tot}}{\partial c_{i\boldsymbol{g}}^*} = \frac{1}{2} g^2 c_{i\boldsymbol{g}} + \int d\boldsymbol{r} e^{-i\boldsymbol{g}\boldsymbol{r}} V_{\rm eff}(\boldsymbol{r}) \phi_i(\boldsymbol{r}), \tag{A.4}$$

which should be calculated in the program code. The real space integration in Eq. (A.4) is numerically done on the mesh grid of the Cartesian coordinates. The

mesh interval for the real space grid is chosen so as to reproduce all the plane waves $\{e^{i\boldsymbol{g}\boldsymbol{r}}\}$ within the cutoff ($|\boldsymbol{g}| < g_c$). In the program code, the Fast Fourier Transform (FFT) algorithms are used for the Fourier transform in Eq. (A.1) and the inversed Fourier transform in Eq. (A.4). In the first-principle molecular dynamics, the FFT routines usually consume the dominant part of the total computational cost.

Fig.A.1 is two snapshots of an *ab initio* molecular dynamic simulation performed with our original program code [125]. The system is α-NaSn with a partial melting phase at 757K $\leq T \leq$ 854K. The simulation cell contains 64 atoms. The tin atoms form $(Sn_4)^{4-}$ tetrahedrons, as in the low-temperature solid phase, while the sodium atoms show diffusive motions. In each snapshot, the charge density of a selected eigen state is drawn. The selected eigen states are one near the Fermi level or near the highest occupied level. An almost spherical charge distribution, labeled 'n', corresponds to a non-bonding (atomic) orbital on a tin atom, whereas a oval charge distribution, labeled 'b', corresponds to a bonding orbital within a $(Sn_4)^{4-}$ tetrahedron. The simulation describes the dynamical bondbreaking and rebonding processes, which is a typical quantum mechanical process.

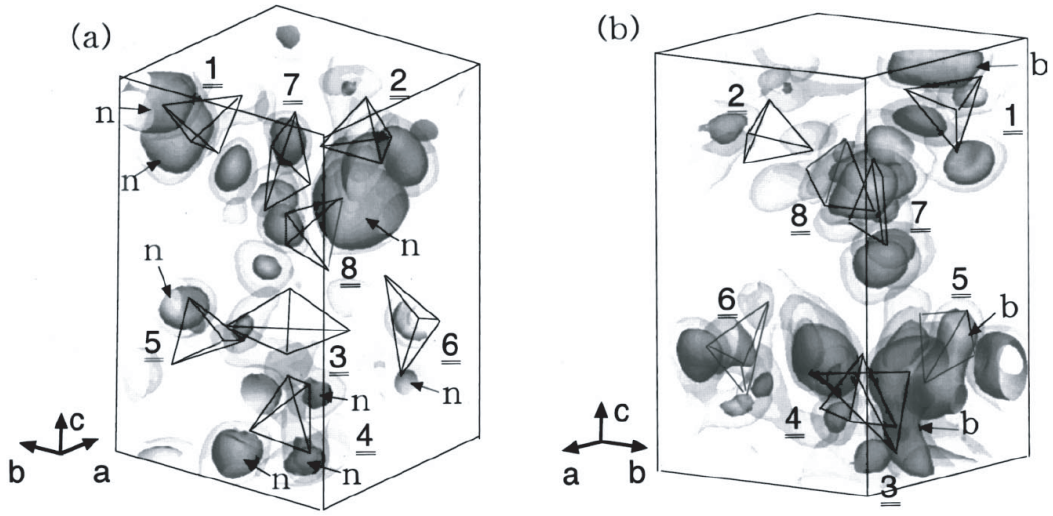


Figure A.1: Snapshots of an *ab initio* molecular dynamics simulation of α-NaSn [125]. In each snapshot, the charge density of a selected eigen state is drawn. The snapshot (b) is one after (a) by the time interval of 0.3 ps. Eight tetragonal $(Sn_4)^{4-}$ are included and are labeled with the numbers $1, 2, 3...8$. Two figures are rotated to show the bonding character clearly. For the charge density, the characteristic bonding or non-bonding regions are indicated 'b' or 'n', respectively.

Limitation of the LDA

The second topic is the limitation of the LDA. Though the LDA reproduces a vast number of experimental results, there are several systematic problems. See reviews, such as Ref. [8]. To overcome such problems, several methods are being developed. They are generally called 'beyond LDA methods'. Now we discuss only one problem of the LDA, which are related to the discussion of this thesis.

Within the LDA, the band gap in semiconductors is usually underestimated. This problem is overcome by the GW approximation (GWA) [126], one of the beyond LDA method. The GW approximation is based on the many body perturbation theory and gives the quasi particle picture. See a review article [127] for details and recent developments. As a typical example, the LDA and GWA results are compared in the band structures of the silicon crystal, which can be seen in several papers, such as Fig.5 of Ref.[128]. Other references can be seen in the review articles [127]. Figure A.2(a) shows the band structures of the silicon, in which the LDA result is plotted by lines and the GWA result by dots. The experimental value of the band gap is 1.17 eV. The LDA result of the band gap is too small (0.60 eV), while the GWA result gives a reasonable value (1.26 eV). In Fig. A.2(b), the LDA result for the conduction band ($\varepsilon > 0$) is shifted *artificially* upward by 0.66 eV, so as to reproduce the band gap of the GWA. A good agreement is seen between the shifted LDA result and the GWA result. In short, the problem of the LDA is solved by an artificial shift of the conduction band. A similar situation can be seen on the Si(001) surface [129], which is reviewed in Ref.[127]; The GWA result of the unoccupied surface state is shifted upward by a constant energy value from the LDA result.

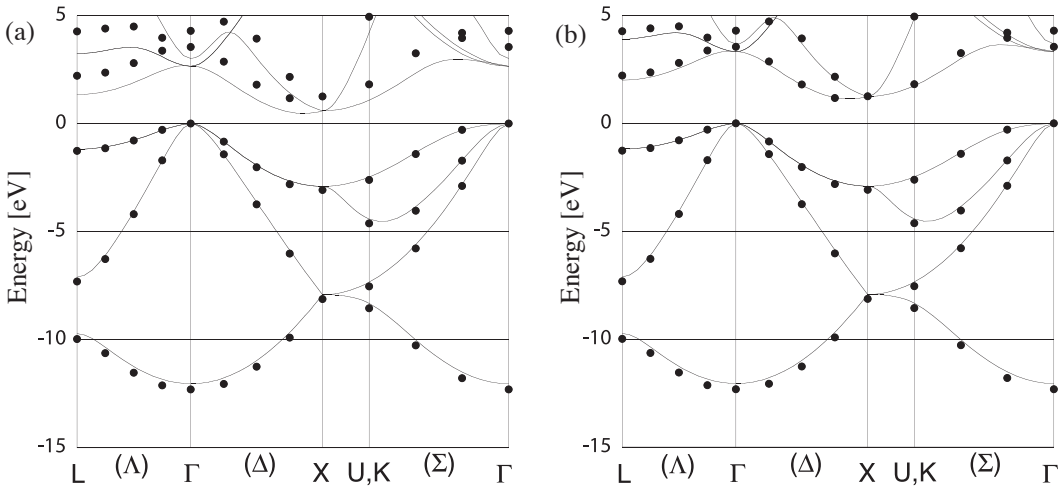


Figure A.2: The comparison of the LDA (line) and GWA (dot) results in the band structure of silicon. The top of the valence band is chosen as the origin of the energy axis. The two figures (a) and (b) differ only in the fact that, in (b), the LDA result of the conduction bands is shifted upward by 0.66 eV. See Ref.[128] or the review article [127], as references. The present figures are plotted based on the data by A. Yamasaki [130].

A.2 Tight-binding formulation

In this appendix, we explain the formulation of practical tight-binding Hamiltonians with the Slater-Koster form[60]. The explanation is done among s and p orbitals with the example of the diamond structure.

Minimal Hamiltonian with s and p orbitals

The minimal tight-binding Hamiltonian is described by the nearest neighbor interaction among the s, p_x, p_y,p_z orbitals. Since the diamond structure contains two atoms in the primitive cell, the nearest neighbor tight-binding Hamiltonian is given as an 8×8 matrix. The explicit parametrization for silicon and the other diamond structure solids was done in many papers, [69, 62]. The minimal Hamiltonian contains the following six parameters, in the Slater-Koster form[60],

$$\varepsilon_{\rm s}, \quad \varepsilon_{\rm p}, \quad V_{\rm ss\sigma}, \quad V_{\rm sp\sigma}, \quad V_{\rm pp\sigma}, \quad V_{\rm pp\pi}. \tag{A.5}$$

Among them, the two parameters $V_{\rm pp\sigma}, V_{\rm pp\pi}$ can be transformed into the following two parameters

$$V_{xx} \equiv \frac{1}{3}V_{\rm pp\sigma} + \frac{2}{3}V_{\rm pp\sigma} \tag{A.6}$$

$$V_{xy} \equiv \frac{1}{3}V_{\rm pp\sigma} - \frac{1}{3}V_{\rm pp\sigma} \tag{A.7}$$

Figure A.3 shows the schematic picture of the hopping integrals of $V_{\rm ss\sigma}, V_{\rm sp\sigma}, V_{\rm pp\sigma}, V_{\rm pp\pi}$.

Here we write down the expressions of matrix elements between the s and p orbital in a general form; We define the s and p orbital on the i-th atom as $|{\rm s}_i\rangle$ and

$$|{\rm p}\boldsymbol{a}_i\rangle \equiv a_{xi}|{\rm p}_{xi}\rangle + a_{yi}|{\rm p}_{yi}\rangle + a_{zi}|{\rm p}_{zi}\rangle, \tag{A.8}$$

respectively. Here the vector $\boldsymbol{a}_i \equiv (a_{xi}, a_{yi}, a_{zi})$ indicates the direction of the p orbital $|{\rm p}\boldsymbol{a}_i\rangle$ ($a_{xi}^2 + a_{yi}^2 + a_{zi}^2 = 1$). The positions of the i-th atoms ($i = 1, 2$) are defined as $\boldsymbol{R}_1, \boldsymbol{R}_2$. The vectors $\boldsymbol{r}$, $\hat{\boldsymbol{r}}$ are defined, respectively, as

$$\boldsymbol{r} \equiv \boldsymbol{R}_2 - \boldsymbol{R}_1, \quad \hat{\boldsymbol{r}} \equiv \frac{\boldsymbol{r}}{|\boldsymbol{r}|}. \tag{A.9}$$

With these definitions, several matrix elements can be written as

$$\langle {\rm s}_1|H|{\rm p}\boldsymbol{a}_2\rangle = (\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})V_{\rm sp\sigma} \tag{A.10}$$

$$\langle {\rm p}\boldsymbol{a}_1|H|{\rm p}\boldsymbol{a}_2\rangle = (\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})(\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})\, V_{\rm pp\sigma} + (\boldsymbol{b}_1 \cdot \boldsymbol{b}_2)\, V_{\rm pp\pi}, \tag{A.11}$$

$$\tag{A.12}$$

where we use the notation

$$\boldsymbol{b}_i \equiv \boldsymbol{a}_i - (\boldsymbol{a}_i \cdot \hat{\boldsymbol{r}})\, \hat{\boldsymbol{r}} \quad (i = 1, 2). \tag{A.13}$$

Figure A.4(a) is the schematic picture of Eq. (A.10). Figure A.4(b) shows the vectors $\hat{\boldsymbol{r}}, \boldsymbol{a}_i, \boldsymbol{b}_i$. Note that the inner product $\boldsymbol{b}_1 \cdot \boldsymbol{b}_2$ in Eq. (A.11) is rewritten as

$$\begin{aligned} \boldsymbol{b}_1 \cdot \boldsymbol{b}_2 &= \{\boldsymbol{a}_1 - (\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})\, \hat{\boldsymbol{r}}\} \cdot \{\boldsymbol{a}_2 - (\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})\, \hat{\boldsymbol{r}}\} \\ &= (\boldsymbol{a}_1 \cdot \boldsymbol{a}_2) - (\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})(\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}}) \end{aligned} \tag{A.14}$$

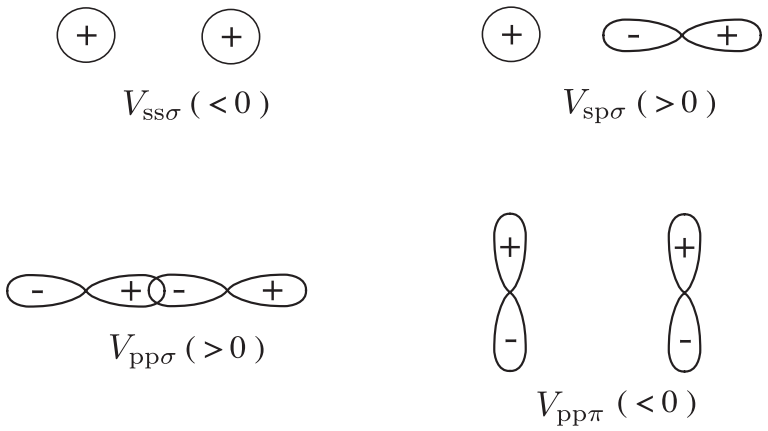


Figure A.3: Schematic picture of the hopping integrals among s and p orbitals.

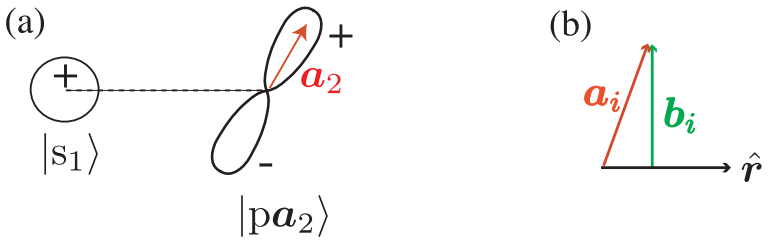


Figure A.4: (a) Schematic picture of the Hamiltonian matrix element $\langle s_1|H|p\boldsymbol{a}_2\rangle$. (b) Figure of the vectors $\hat{\boldsymbol{r}}, \boldsymbol{a}_i, \boldsymbol{b}_i$.

Hamiltonian with the extra 's*' orbital

An improved description of the electronic structure is given by the formulation with the extra 's*' orbital [61], which was discussed in Section 3.1. The extra s* orbital is in the s (spherical) symmetry and its physical origin could be the spherical average of the five empty d orbitals. The resultant Hamiltonian is given as a 10×10 matrix, which contains two additional parameters;

$$\varepsilon_{s^*}, \quad V_{s^*p\sigma}. \tag{A.15}$$

In other words, the parameter $V_{s^*p\sigma}$ corresponds to the average of the d-p interaction. The formulation will be reduced to the minimal formulation with $V_{s^*p\sigma} = 0$. The practical values of these parameters are shown in Table A.1 for the diamond structure solids. Note that the original paper [61] also contains the parameters in zincblend structure solids. The resultant band structure for silicon is shown in Fig.A.5(a), in which the unit cell for the Brillouin zone is cubic as commonly used [62, 131]. The corresponding *ab initio* calculation can be seen in Fig.A.2(b). Figure A.5(b) shows the result of the modified Hamiltonian in which $V_{sp\sigma} = V_{s^*p\sigma} = 0$ is used and the other parameters are not modified from those in silicon. In Fig.A.5(b), the s, p and s* bands are decoupled and, especially, the s* band is dispersion-less.

The formulations are not explained in detail. Instead, we discuss the eigen levels at several points in the Brillouin zone. We pick out the Γ point and the X point, because the Hamiltonian matrix is solved analytically at these points. In silicon,

	C	Si	Ge	α-Sn
$\varepsilon_{\rm s}$	-4.5450	-4.2000	-5.8800	-5.6700
$\varepsilon_{\rm p}$	3.8400	1.7150	1.6100	1.3300
$\varepsilon_{\rm s^*}$	11.3700	6.6850	6.3900	5.9000
$4V_{\rm sp\sigma}$	15.2206	5.7292	5.4649	4.5116
$4V_{\rm s^*p\sigma}$	8.2109	5.3749	5.2191	5.8939
$4V_{\rm ss}$	-22.7250	-8.3000	-6.7800	-5.6700
$4V_{xx}$	3.8400	1.7150	1.6100	1.3300
$4V_{xy}$	11.6700	4.5750	4.9000	4.0800
$\varepsilon_{\rm p}-\varepsilon_{\rm s}$	8.3850	5.9150	7.4900	7.0000
$\varepsilon_{\rm s^*}-\varepsilon_{\rm p}$	7.5300	4.9700	4.7800	4.5700

Table A.1: The energy parameters of the tight-binding Hamiltonian in Ref. [61]. The parameters are in the unit of eV.

the top and bottom of the valence band appears at the Γ point and the bottom of the conduction band appear near the X point in Fig. A.5(a).

At the Γ point, the eigen levels ε are decomposed into the s, p and s* bands as follows;

$$\varepsilon = \varepsilon_{\rm s} - 4V_{\rm ss\sigma}, \quad n=1 \tag{A.16}$$
$$\varepsilon = \varepsilon_{\rm s} + 4V_{\rm ss\sigma}, \quad n=1 \tag{A.17}$$
$$\varepsilon = \varepsilon_{\rm p} - 4V_{xx}, \quad n=3 \tag{A.18}$$
$$\varepsilon = \varepsilon_{\rm p} + 4V_{xx}, \quad n=3 \tag{A.19}$$
$$\varepsilon = \varepsilon_{\rm s^*}, \quad n=2 \tag{A.20}$$

Here and hereafter, n is the number of the degeneracy of each level. Since the eigen levels are independent on $V_{\rm sp\sigma}$ and $V_{\rm s^*p\sigma}$, the eigen levels at the the Γ point are the same between Fig. A.5(a) and (b).

At the X point, the Hamiltonian matrix is decomposed into four block matrices; Two of them are the same 2×2 matrix, which gives the four eigen levels

$$\varepsilon = \varepsilon_{\rm p} - 4V_{xy}, \quad n=2 \tag{A.21}$$
$$\varepsilon = \varepsilon_{\rm p} + 4V_{xy}, \quad n=2. \tag{A.22}$$

These levels appear commonly in Fig.A.5(a) and (b). The other two block matrices are the same the 3×3 matrix

$$\begin{pmatrix} \varepsilon_{\rm s} & -4iV_{\rm sp\sigma} & 0 \\ 4iV_{\rm sp\sigma} & \varepsilon_{\rm p} & -4iV_{\rm s^*p\sigma} \\ 0 & 4iV_{\rm s^*p\sigma} & \varepsilon_{\rm s^*} \end{pmatrix} \tag{A.23}$$

and, therefore, the three eigen levels are doubly degenerated ($n=2$). From Eq. (A.23), the three eigen levels are equal to the atomic levels, $\varepsilon_{\rm s}, \varepsilon_{\rm p}, \varepsilon_{\rm s^*}$, at the X point in Fig. A.5 (b) , because of $V_{\rm sp\sigma} = V_{\rm s^*p\sigma} = 0$. Moreover, if one chooses $\varepsilon_{\rm s^*}$ and $V_{\rm s^*p}$ so as to

$$\varepsilon_{\rm s^*} - \varepsilon_{\rm p} = \varepsilon_{\rm p} - \varepsilon_{\rm s} \tag{A.24}$$
$$V_{\rm s^*p} = V_{\rm sp}, \tag{A.25}$$

the eigen levels are easily obtained as

$$\varepsilon = \varepsilon_{\rm p} - \sqrt{(\varepsilon_{\rm p} - \varepsilon_{\rm s})^2 + 2\,(4V_{\rm sp})^2}, \quad n = 2 \tag{A.26}$$

$$\varepsilon = \varepsilon_{\rm p}, \quad n = 2 \tag{A.27}$$

$$\varepsilon = \varepsilon_{\rm p} + \sqrt{(\varepsilon_{\rm p} - \varepsilon_{\rm s})^2 + 2\,(4V_{\rm sp})^2}, \quad n = 2. \tag{A.28}$$

In Table A.1, we can find that the above choices in Eqs. (A.24),(A.25) are roughly coincident with the silicon case.

Finally, the difference of the eigen levels at the Γ and X points is discussed between the present formulation, with s, p_x,p_y,p_z,s^* orbitals, and the minimal formulation, with s, p_x,p_y,p_z orbitals. No difference is seen among the eigen levels given by Eqs. (A.16), (A.17),(A.18),(A.19), (A.21),(A.22). The essential difference is the fact that the 3×3 matrix in Eq. (A.23) will be reduced to the 2×2 matrix

$$\begin{pmatrix} \varepsilon_{\rm s} & -4iV_{\rm sp} \\ 4iV_{\rm sp} & \varepsilon_{\rm p} \end{pmatrix}, \tag{A.29}$$

in the minimal formulation.

Mathematical notes

Here several useful mathematical relations are added; A useful relation is given as

$$\frac{\partial}{\partial \boldsymbol{r}}\,(\boldsymbol{a}_i \cdot \hat{\boldsymbol{r}}) = -\frac{1}{r}\boldsymbol{b}_i \quad (i = 1, 2). \tag{A.30}$$

In practical tight-binding Hamiltonians for molecular dynamics, the hopping integrals $V_{{\rm ss}\sigma}, V_{{\rm sp}\sigma}, V_{{\rm pp}\sigma}, V_{{\rm pp}\pi}$ are the functions of the interatomic distance $r \equiv |\boldsymbol{r}|$. Using Eq. (A.30), we write down the gradients of several terms in Eq. (A.10) and (A.11) as follows;

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{r}}\{(\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})V_{{\rm sp}\sigma}(r)\} \\ = \;& (\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})V'_{{\rm sp}\sigma}(r)\hat{\boldsymbol{r}} - \frac{1}{r}V_{{\rm sp}\sigma}(r)\boldsymbol{b}_2 \end{aligned} \tag{A.31}$$

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{r}}\left\{(\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})(\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})\,V_{{\rm pp}\sigma}(r)\right\} \\ = \;& (\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})(\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})V'_{{\rm pp}\sigma}(r)\hat{\boldsymbol{r}} \\ - \;& (\boldsymbol{a}_1 \cdot \hat{\boldsymbol{r}})V_{{\rm pp}\sigma}(r)\frac{1}{r}\boldsymbol{b}_2 - (\boldsymbol{a}_2 \cdot \hat{\boldsymbol{r}})V_{{\rm pp}\sigma}(r)\frac{1}{r}\boldsymbol{b}_1. \end{aligned} \tag{A.32}$$

These gradients are calculated in the program code, ao as to calculate the force on an atom

$$\frac{\partial E_{\rm elec}}{\partial \boldsymbol{R}_I}. \tag{A.33}$$

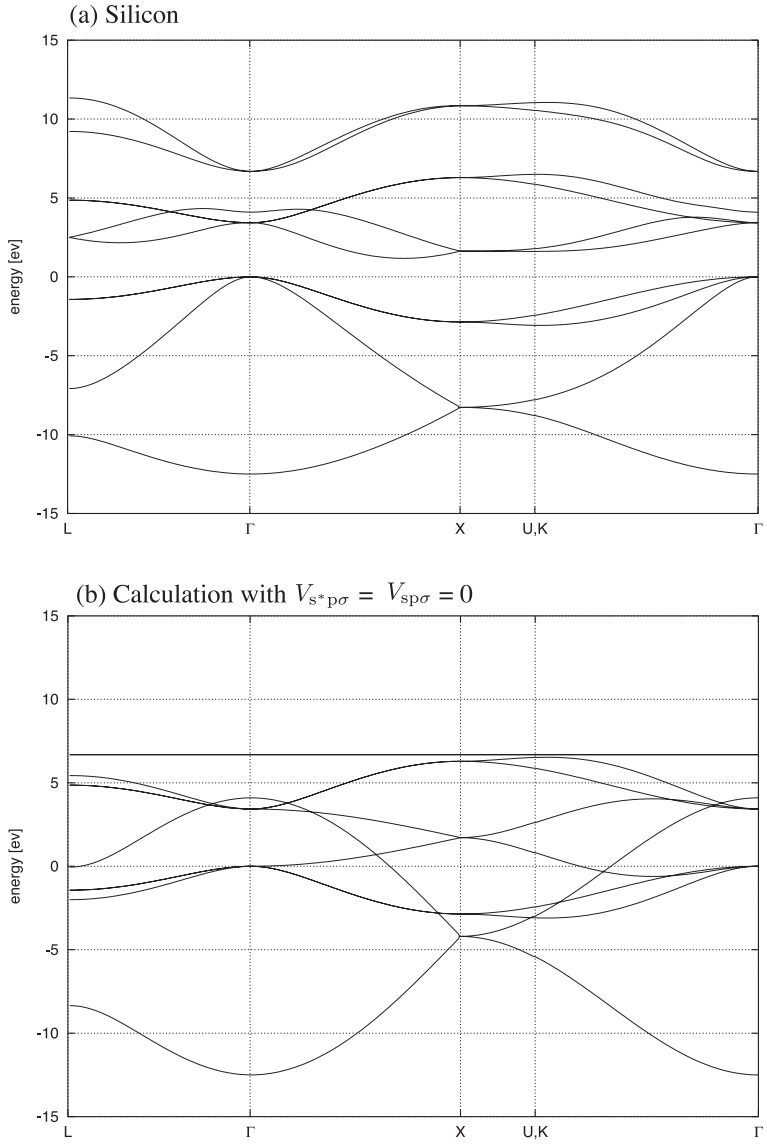


Figure A.5: Band structure of silicon using a tight-binding Hamiltonian with the extra 's∗' orbital [61]; No parameter is modified for (a), while the parameters are modified in $V_{sp\sigma} = V_{s^*p\sigma} = 0$ for (b).

Appendix B

Theory of elasticity

B.1 Theory in cubic symmetry

This appendix is devoted to the elastic theory within the cubic symmetry. See Refs. [131, 132], as standard textbooks.

Definitions of strain tensors

Consider nearly located two points $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Under a deformation substantially uniform near $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, the vector $\delta\boldsymbol{r} \equiv \boldsymbol{r}_1$ - $\boldsymbol{r}_2$ is changed into

$$\delta\boldsymbol{r} \rightarrow \delta\boldsymbol{r}' \equiv \delta\boldsymbol{r} + \delta\boldsymbol{u}. \tag{B.1}$$

This is the definition of the displacement vector $\delta\boldsymbol{u}$. Now we introduce the unit vector of the Cartesian coordinates $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3) \equiv (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ satisfying $\hat{\boldsymbol{x}}_i \cdot \hat{\boldsymbol{x}}_j = \delta_{ij}$. The vectors $\delta\boldsymbol{r}$, $\delta\boldsymbol{r}'$ and $\delta\boldsymbol{u}$ are written, respectively, by

$$\begin{aligned}
\delta\boldsymbol{r} &\equiv \sum_{i=1}^{3}(\delta x_i)\hat{\boldsymbol{x}}_i \\
\delta\boldsymbol{r}' &\equiv \sum_{i=1}^{3}(\delta x_i')\hat{\boldsymbol{x}}_i \\
\delta\boldsymbol{u} &\equiv \sum_{i=1}^{3}(\delta u_i)\hat{\boldsymbol{x}}_i.
\end{aligned}$$

Here the vector $\delta\boldsymbol{u}$ is the function of $(\delta x, \delta y, \delta z)$, which will be zero in the case of $(\delta x, \delta y, \delta z) = (0, 0, 0)$. Within the linear-order expansion of δu_j

$$\delta u_j \approx \sum_i \frac{\partial u_j}{\partial x_i}\delta x_i, \tag{B.2}$$

we can write

$$\begin{aligned}
\delta\boldsymbol{r}' &\equiv \delta\boldsymbol{r} + \delta\boldsymbol{u} \\
&= \sum_i(\delta x_i)\hat{\boldsymbol{x}}_i + \sum_j(\delta u_j)\hat{\boldsymbol{x}}_j \\
&= \sum_i(\delta x_i)\hat{\boldsymbol{x}}_i + \sum_{ij}\frac{\partial u_j}{\partial x_i}(\delta x_i)\hat{\boldsymbol{x}}_j \\
&= \sum_i(\delta x_i)\left\{\hat{\boldsymbol{x}}_i + \sum_j\frac{\partial u_j}{\partial x_i}\hat{\boldsymbol{x}}_j\right\}.
\end{aligned} \tag{B.3}$$

Equation (B.3) can be rewritten as

$$\delta\boldsymbol{r}' = \sum_i(\delta x_i)\hat{\boldsymbol{x}}_i' \tag{B.4}$$

with the definition of

$$\begin{aligned}
\hat{\boldsymbol{x}}_i' &\equiv \hat{\boldsymbol{x}}_i + \sum_j\frac{\partial u_j}{\partial x_i}\hat{\boldsymbol{x}}_j \\
&= \left(1 + \frac{\partial u_i}{\partial x_i}\right)\hat{\boldsymbol{x}}_i + \sum_{j(\neq i)}\frac{\partial u_j}{\partial x_i}\hat{\boldsymbol{x}}_j.
\end{aligned} \tag{B.5}$$

Now the deformation ($\delta \boldsymbol{r} \to \delta \boldsymbol{r}'$) is interpreted as the deformation of the axis vector ($\hat{\boldsymbol{x}}_i \to \hat{\boldsymbol{x}'}_i$), where the components $(\delta x, \delta y, \delta z)$ are unchanged ($\sum_i (\delta x_i)\hat{\boldsymbol{x}}_i \to \sum_i (\delta x_i)\hat{\boldsymbol{x}}'_i$).

The components of the strain tensor are defined as, for example,

$$\begin{aligned} e_{xx} &\equiv \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{x}}' - 1 \\ e_{xy} &\equiv \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}}' (= e_{yx}), \end{aligned} \tag{B.6}$$

which are zero without any deformation. All the independent components are $(e_{xx}, e_{yy}, e_{zz}, e_{xy}, e_{yz}, e_{zx})$. When we substitute Eq.(B.5) into Eq.(B.6) and ignore the second order of $(\partial u/\partial x)$, we obtain

$$\begin{aligned} e_{xx} &= \frac{\partial u_1}{\partial x_1} \\ e_{xy} &= \frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2} = e_{yx}. \end{aligned} \tag{B.7}$$

Definitions of stress tensors

In general, the stress tensor σ has nine components :

$$\sigma \equiv \begin{pmatrix} X_x & X_y & X_z \\ Y_x & Y_y & Y_y \\ Z_x & Z_y & Z_z \end{pmatrix}. \tag{B.8}$$

The capital letters indicate the direction of the force and the subscript indicates the normal vector of the plane on which the stress is imposed. The stress on a plane of which normal vector is $\vec{n}$ is given by

$$\sigma \vec{n}. \tag{B.9}$$

For example, (X_x, Y_x, Z_x) is the stress on the yz-plane. Now the theory is restricted to the systems without rotational forces

$$X_y = Y_x, \quad Y_z = Z_y, \quad Z_x = X_z. \tag{B.10}$$

The number of the independent components in the stress tensor is reduced from nine to six and the stress tensor σ is reduced to a symmetric matrix.

Strain energy and stiffness constants

In general, the strain energy is written within the second order of the components of strain. With the cubic symmetry, the strain energy is reduced to

$$\begin{aligned} U &= \frac{C_{11}}{2}\left(e_{xx}^2 + e_{yy}^2 + e_{zz}^2\right) + C_{12}\left(e_{xx}e_{yy} + e_{yy}e_{zz} + e_{zz}e_{xx}\right) \\ &+ \frac{C_{44}}{2}\left(e_{xy}^2 + e_{yz}^2 + e_{zx}^2\right), \end{aligned} \tag{B.11}$$

where the three independent parameters (C_{11}, C_{12}, C_{44}) are called 'elastic constants' or 'stiffness constants'. They characterize the elastic property of materials.

The components of the stress tensor are given by, for examples,

$$X_x = \frac{\partial U}{\partial e_{xx}} = C_{11}e_{xx} + C_{12}\left(e_{yy} + e_{zz}\right) \tag{B.12}$$

$$X_y = \frac{\partial U}{\partial e_{xy}} = C_{44}e_{xy}. \tag{B.13}$$

In a matrix formula, we can write

$$\begin{pmatrix} X_x \\ Y_y \\ Z_z \\ X_y \\ Y_z \\ Z_x \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{12} & & & \\ C_{12} & C_{11} & C_{12} & & & \\ C_{12} & C_{12} & C_{11} & & & \\ & & & C_{44} & & \\ & & & & C_{44} & \\ & & & & & C_{44} \end{pmatrix} \begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \\ e_{xy} \\ e_{yz} \\ e_{zx} \end{pmatrix}. \tag{B.14}$$

When we define the smaller matrix $\bar{C}$ as

$$\bar{C} \equiv \begin{pmatrix} C_{11} & C_{12} & C_{12} \\ C_{12} & C_{11} & C_{12} \\ C_{12} & C_{12} & C_{11} \end{pmatrix}, \tag{B.15}$$

its eigen vectors are obtained as

$$\begin{pmatrix} C_{11} & C_{12} & C_{12} \\ C_{12} & C_{11} & C_{12} \\ C_{12} & C_{12} & C_{11} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = (C_{11} + 2C_{12}) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{B.16}$$

$$\begin{pmatrix} C_{11} & C_{12} & C_{12} \\ C_{12} & C_{11} & C_{12} \\ C_{12} & C_{12} & C_{11} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = (C_{11} - C_{12}) \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \tag{B.17}$$

$$\begin{pmatrix} C_{11} & C_{12} & C_{12} \\ C_{12} & C_{11} & C_{12} \\ C_{12} & C_{12} & C_{11} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = (C_{11} - C_{12}) \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}. \tag{B.18}$$

Elastic modes

From the above eigen value analysis, the six freedoms $\{e_{xx}, e_{yy}, e_{zz}, e_{xy}, e_{yz}, e_{zx}\}$ are classified by the following three elastic modes, which are illustrated in Fig. B.1.

(a) Volume expansion mode with the bulk modulus B: $e_{xx} = e_{yy} = e_{zz} \equiv \varepsilon \equiv \delta/3$

$$U = \frac{B}{2}\delta^2 \tag{B.19}$$

$$B \equiv \frac{C_{11} + 2C_{12}}{3} \tag{B.20}$$

Here δ is the ratio of the volume expansion.

(b) Shear mode with the shear modulus $C_{11} - C_{12}$: $e_{xx} = -e_{yy} \equiv \varepsilon$

$$U = (C_{11} - C_{12})\varepsilon^2. \tag{B.21}$$

(c) Shear mode with the shear modulus C_{44} : $e_{xy} \equiv \theta$

$$U = \frac{C_{44}}{2}\theta^2. \tag{B.22}$$

The number of degeneracy is one for (a), two for (b) and three for (c). As we see in the end of this appendix, the isotropic property requires the relation

$$C_{44} = \frac{C_{11} - C_{12}}{2}, \tag{B.23}$$

in which the two shear modes, (b) and (c), will be degenerated.

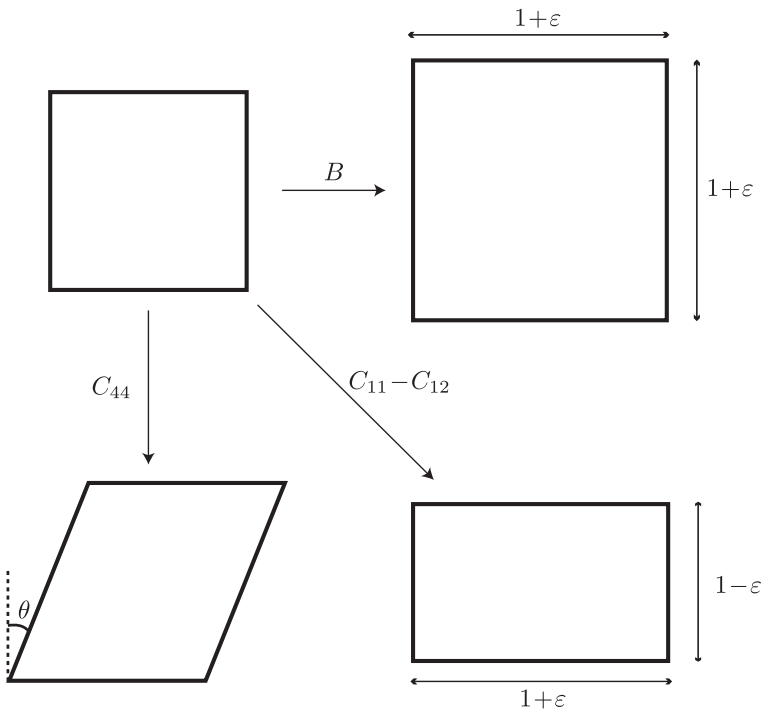


Figure B.1: Schematic figure of the three elastic modes in the cubic symmetry. The elastic modes are labeled by the corresponding elastic constants: the bulk modulus (B) and the two shear moduli ($C_{11} - C_{12}$, C_{44}).

Compliance constants

When we solve Eq. (B.14) with respect to the components of strain, we obtain

$$\begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \\ e_{xy} \\ e_{yz} \\ e_{zx} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{12} & & & \\ C_{12} & C_{11} & C_{12} & & & \\ C_{12} & C_{12} & C_{11} & & & \\ & & & C_{44} & & \\ & & & & C_{44} & \\ & & & & & C_{44} \end{pmatrix}^{-1} \begin{pmatrix} X_x \\ Y_y \\ Z_z \\ X_y \\ Y_z \\ Z_x \end{pmatrix}$$

$$= \begin{pmatrix} S_{11} & S_{12} & S_{12} & & & \\ S_{12} & S_{11} & S_{12} & & & \\ S_{12} & S_{12} & S_{11} & & & \\ & & & S_{44} & & \\ & & & & S_{44} & \\ & & & & & S_{44} \end{pmatrix} \begin{pmatrix} X_x \\ Y_y \\ Z_z \\ X_y \\ Y_z \\ Z_x \end{pmatrix}, \tag{B.24}$$

where

$$\det[\bar{C}] = (C_{11} + 2C_{12})(C_{11} - C_{12})^2 \tag{B.25}$$

$$S_{44} = \frac{1}{C_{44}} \tag{B.26}$$

$$\begin{aligned} S_{11} &= \left(\bar{C}^{-1}\right)_{11} = \frac{1}{\det[\bar{C}]} \begin{vmatrix} C_{11} & C_{12} \\ C_{12} & C_{11} \end{vmatrix} \\ &= \frac{C_{11}^2 - C_{12}^2}{(C_{11} + 2C_{12})(C_{11} - C_{12})^2} \\ &= \frac{C_{11} + C_{12}}{(C_{11} + 2C_{12})(C_{11} - C_{12})} \end{aligned} \tag{B.27}$$

$$\begin{aligned} S_{12} &= \left(\bar{C}^{-1}\right)_{12} = \frac{-1}{\det[\bar{C}]} \begin{vmatrix} C_{12} & C_{12} \\ C_{12} & C_{11} \end{vmatrix} \\ &= -\frac{C_{12}(C_{11} - C_{12})}{(C_{11} + 2C_{12})(C_{11} - C_{12})^2} \\ &= \frac{-C_{12}}{(C_{11} + 2C_{12})(C_{11} - C_{12})}. \end{aligned} \tag{B.28}$$

The three independent parameters (S_{11}, S_{12}, S_{44}) and called 'compliance constants', which are equivalent to the stiffness constants (C_{11}, C_{12}, C_{44}). Also note that

$$S_{11} - S_{12} = \frac{1}{C_{11} - C_{12}}. \tag{B.29}$$

Young modulus at (100) direction

As an example of the anisotropic elastic properties, the Young modulus and the Poisson ratio are calculated under the external load in the (100) direction $(E_{100}, \nu_{100}^{(0)})$. The present Poisson ratio $\nu_{100}^{(0)}$ is defined for the deformation in the (010) or (001) direction, which gives

$$e_{yy} = e_{zz} = -\nu_{100}^{(0)} e_{xx}. \tag{B.30}$$

The components of the stress tensor are zero except X_x. The non-zero components of the strain tensor (e_{xx}, e_{yy}, e_{zz}) are determined by

$$E_{100}e_{xx} = X_x = \frac{\partial U}{\partial e_{xx}} = C_{11}e_{xx} + C_{12}\left(e_{yy} + e_{zz}\right) \tag{B.31}$$

$$0 = Y_y = \frac{\partial U}{\partial e_{yy}} = C_{11}e_{yy} + C_{12}\left(e_{zz} + e_{xx}\right) \tag{B.32}$$

$$0 = Z_z = \frac{\partial U}{\partial e_{zz}} = C_{11}e_{zz} + C_{12}\left(e_{xx} + e_{yy}\right). \tag{B.33}$$

When Eq. (B.30) is introduced into (B.32) or (B.33), we obtain

$$-C_{11}\nu_{100}^{(0)} + C_{12}(-\nu_{100}^{(0)} + 1) = 0 \tag{B.34}$$

and thus

$$\nu_{100}^{(0)} = \frac{C_{12}}{C_{11} + C_{12}}. \tag{B.35}$$

The above equation and Eq. (B.31) give the Young modulus as

$$\begin{aligned} E_{100} &= C_{11} + C_{12}(-2\nu_{100}^{(0)}) \\ &= \frac{C_{11}^2 + C_{11}C_{12} - 2C_{12}^2}{C_{11} + C_{12}} \\ &= \frac{(C_{11} - C_{12})(C_{11} + 2C_{12})}{C_{11} + C_{12}}. \end{aligned} \tag{B.36}$$

The same expression can be obtained from the compliance constants, when the matrix equation

$$\begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{12} \\ S_{12} & S_{11} & S_{12} \\ S_{12} & S_{12} & S_{11} \end{pmatrix} \begin{pmatrix} X_x \\ Y_y \\ Z_z \end{pmatrix} \tag{B.37}$$

is reduced to

$$\begin{pmatrix} 1 \\ -\nu_{100}^{(0)} \\ -\nu_{100}^{(0)} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{12} \\ S_{12} & S_{11} & S_{12} \\ S_{12} & S_{12} & S_{11} \end{pmatrix} \begin{pmatrix} E_{100} \\ 0 \\ 0 \end{pmatrix}. \tag{B.38}$$

Using Eqs. (B.27) and (B.28), the Young modulus and Poisson ratio are given, as

$$E_{100} = \frac{1}{S_{11}} = \frac{(C_{11} - C_{12})(C_{11} + 2C_{12})}{C_{11} + C_{12}} \tag{B.39}$$

$$\nu_{100}^{(0)} = -E_{100}S_{12} = -\frac{S_{12}}{S_{11}} = \frac{C_{12}}{C_{11} + C_{12}}. \tag{B.40}$$

Another definition of strain tensor

Now we introduce another definition of the strain tensor, which are seen in some textbooks. The nine variables u_{ij} are defined as

$$\begin{aligned} u_{xx} &= \frac{\partial u_1}{\partial x_1} = e_{xx} \\ u_{xy} &= \frac{1}{2}\left(\frac{\partial u_y}{\partial x} + \frac{\partial u_x}{\partial y}\right)\left(= \frac{1}{2}e_{yx}\right) \\ u_{yx} &= \frac{1}{2}\left(\frac{\partial u_y}{\partial x} + \frac{\partial u_x}{\partial y}\right)\left(= \frac{1}{2}e_{yx}\right) \end{aligned} \tag{B.41}$$

and so on. The strain energy is defined as a function of the *nine* independent variables;

$$\begin{aligned} U &= U(u_{xx}, u_{yy}, u_{zz}, u_{xy}, u_{yx}, u_{yz}, u_{zy}, u_{zx}, u_{xz}) \\ &= \frac{C_{11}}{2}\left(u_{xx}^2 + u_{yy}^2 + u_{zz}^2\right) + C_{12}\left(u_{xx}u_{yy} + u_{yy}u_{zz} + u_{zz}u_{xx}\right) \\ &+ 2C_{44}\left(u_{xy}u_{yx} + u_{yz}u_{zy} + u_{zx}u_{zx}\right). \end{aligned} \tag{B.42}$$

The stress tensor is redefined as

$$\sigma_{ij} \equiv \left(\frac{\partial U}{\partial u_{ij}}\right)_{9\text{ variables}}, \tag{B.43}$$

where U is differentiated as the function of the *nine* independent variables. In results, we obtain the matrix representation of the stress tensor σ, such as

$$\sigma_{xx} \equiv \left(\frac{\partial U}{\partial u_{xx}}\right)_{9\text{ variables}} = C_{11}u_{xx} + C_{12}(u_{yy} + u_{zz}) \tag{B.44}$$

$$\sigma_{xy} \equiv \left(\frac{\partial U}{\partial u_{xy}}\right)_{9\text{ variables}} = 2C_{44}u_{yx}. \tag{B.45}$$

The expressions with e_{ij} can be obtained with the relation

$$\frac{\partial U}{\partial e_{ij}} = \frac{1}{2}\left\{\left(\frac{\partial U}{\partial u_{ij}}\right)_{9\text{ variables}} + \left(\frac{\partial U}{\partial u_{ji}}\right)_{9\text{ variables}}\right\}. \tag{B.46}$$

The formulation with u_{ij} is useful in several mathematical formulations. For instance, the energy description of Eq. B.42 is equivalent to

$$U = \sum_{ij}\sigma_{ij}u_{ij} = \mathrm{Tr}[\sigma u] \tag{B.47}$$

Young modulus at arbitrary direction

Hereafter we derive the Young modulus at an arbitrary direction [132]. With the notations given in this note, the extension ratio is defined given by

$$\varepsilon \equiv \frac{|\delta \boldsymbol{r}'| - |\delta \boldsymbol{r}|}{|\delta \boldsymbol{r}|}, \tag{B.48}$$

which is essential to the calculations of the Young modulus and the Poisson ratio. The linear expansion (B.2) gives

$$\begin{aligned}
|\delta \boldsymbol{r}'|^2 &= \sum_i (\delta x_i')^2 \\
&= \sum_i (\delta x_i + \delta u_i)^2 \\
&= \sum_i (\delta x_i)^2 + \sum_i (\delta x_i)(\delta u_i) + \sum_j (\delta x_j)(\delta u_j) + \sum_k (\delta u_k)(\delta u_k) \\
&= |\delta \boldsymbol{r}|^2 + \sum_{ij} (\delta x_i) \frac{\partial u_i}{\partial x_j} (\delta x_j) + \sum_{ij} (\delta x_j) \frac{\partial u_j}{\partial x_i} (\delta x_i) \\
&\quad + \sum_{ijk} (\delta x_i) \frac{\partial u_k}{\partial x_i} \frac{\partial u_k}{\partial x_j} (\delta x_j)
\end{aligned} \tag{B.49}$$

Here we, again, ignore the second order of $(\partial u/\partial x)$, and obtain

$$|\delta \boldsymbol{r}'|^2 \approx |\delta \boldsymbol{r}|^2 + 2 \sum_{ij} U_{ij} (\delta x_i)(\delta x_j) \tag{B.50}$$

where

$$U_{ij} \equiv \frac{1}{2} \left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j} \right). \tag{B.51}$$

The matrix U_{ij} is equal to the strain matrix e_{ij}, except a factor 1/2 in the off-diagonal elements ($U_{xx} = e_{xx}, U_{xy} = e_{xy}/2$). Using the above relations, $|\delta \boldsymbol{r}'|$ is given by

$$\begin{aligned}
|\delta \boldsymbol{r}'| &\approx \left\{ |\delta \boldsymbol{r}|^2 + 2 \sum_{ij} U_{ij} (\delta x_i)(\delta x_j) \right\}^{1/2} \\
&\approx |\delta \boldsymbol{r}| \left\{ 1 + \sum_{ij} U_{ij} \frac{\delta x_i}{|\delta \boldsymbol{r}|} \frac{\delta x_j}{|\delta \boldsymbol{r}|} \right\}
\end{aligned} \tag{B.52}$$

If the direction vector $\boldsymbol{n} \equiv (n_x, n_y, n_z)$ is defined by

$$n_x \equiv \frac{\delta x}{|\delta \boldsymbol{r}|}, \quad n_y \equiv \frac{\delta y}{|\delta \boldsymbol{r}|}, \quad n_z \equiv \frac{\delta z}{|\delta \boldsymbol{r}|}. \tag{B.53}$$

The extension ratio in the $\boldsymbol{n}$- direction is given by

$$\begin{aligned}
\varepsilon(\boldsymbol{n}) &\equiv \frac{|\delta \boldsymbol{r}'| - |\delta \boldsymbol{r}|}{|\delta \boldsymbol{r}|} \approx \sum_{ij} U_{ij} \frac{\delta x_i}{|\delta \boldsymbol{r}|} \frac{\delta x_j}{|\delta \boldsymbol{r}|} = \sum_{ij} U_{ij} n_i n_j \\
&= e_{xx} n_x^2 + e_{yy} n_y^2 + e_{zz} n_z^2 + e_{xy} n_x n_y + e_{yz} n_y n_z + e_{zx} n_z n_x
\end{aligned} \tag{B.54}$$

Now we turn to consider the situation that a cubic crystal is subject to an external load in the $\boldsymbol{l}$- direction and the Poisson ratio is measured in the $\boldsymbol{m}$- direction. The unit vectors $\boldsymbol{l} \equiv (l_x, l_y, l_z)$ and $\boldsymbol{m} \equiv (m_x, m_y, m_z)$ are orthogonal ($\boldsymbol{l} \cdot \boldsymbol{m} = 0$). The extension ratios in the $\boldsymbol{l}$- and $\boldsymbol{m}$- directions are denoted as $\varepsilon(\boldsymbol{l})$ and $\varepsilon(\boldsymbol{m})$,

respectively. The norm of the tension is denoted as P. The corresponding Young modulus and Poison ratio are defined as

$$E(\boldsymbol{l}) = \frac{P}{\varepsilon(\boldsymbol{l})}$$
$$\nu(\boldsymbol{l}, \boldsymbol{m}) \equiv -\frac{\varepsilon(\boldsymbol{m})}{\varepsilon(\boldsymbol{l})},$$

respectively. For the present situation, the stress tensor σ must has the properties

$$\sigma \boldsymbol{l} = P\boldsymbol{l} \tag{B.55}$$
$$\sigma \boldsymbol{m} = 0 \tag{B.56}$$

This is fulfilled in the form of $\sigma_{ij} = P l_i l_j$ or

$$\begin{pmatrix} X_x & X_y & X_z \\ Y_x & Y_y & Y_y \\ Z_x & Z_y & Z_z \end{pmatrix} = P \begin{pmatrix} l_x l_x & l_x l_y & l_x l_z \\ l_y l_x & l_y l_y & l_y l_z \\ l_z l_x & l_z l_y & l_z l_z \end{pmatrix} \tag{B.57}$$

The diagonal and off-diagonal components are written in, for examples,

$$X_x = P l_x l_x = C_{11} e_{xx} + C_{12} \left(e_{yy} + e_{zz}\right) \tag{B.58}$$
$$X_y = P l_x l_y = C_{44} e_{xy}. \tag{B.59}$$

The off-diagonal component of the strain tensor are given by, for example,

$$e_{xy} = P \frac{l_x l_y}{C_{44}} = P S_{44} l_x l_y \tag{B.60}$$

The diagonal components $\{e_{xx}, e_{yy}, e_{zz}\}$ are given by the matrix equation

$$\begin{pmatrix} C_{11} & C_{12} & C_{12} \\ C_{12} & C_{11} & C_{12} \\ C_{12} & C_{12} & C_{11} \end{pmatrix} \begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \end{pmatrix} = P \begin{pmatrix} l_x^2 \\ l_y^2 \\ l_z^2 \end{pmatrix} \tag{B.61}$$

or

$$\begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \end{pmatrix} = P \begin{pmatrix} S_{11} & S_{12} & S_{12} \\ S_{12} & S_{11} & S_{12} \\ S_{12} & S_{12} & S_{11} \end{pmatrix} \begin{pmatrix} l_x^2 \\ l_y^2 \\ l_z^2 \end{pmatrix} \tag{B.62}$$

The solutions are, for example,

$$\begin{aligned} \frac{e_{xx}}{P} &= S_{11} l_x^2 + S_{12}(l_y^2 + l_z^2) \\ &= S_{11} l_x^2 + S_{12}(1 - l_x^2) \\ &= (S_{11} - S_{12}) l_x^2 + S_{12} \end{aligned} \tag{B.63}$$

or, with the stiffness constants,

$$\frac{e_{xx}}{P} = \frac{1}{C_{11} - C_{12}} l_x^2 - \frac{C_{12}}{(C_{11} + 2C_{12})(C_{11} - C_{12})}. \tag{B.64}$$

With the definition

$$\gamma_1(\boldsymbol{l}) \equiv l_x^2 l_y^2 + l_y^2 l_z^2 + l_z^2 l_x^2 \tag{B.65}$$

and the relation

$$1 = (\boldsymbol{l}\cdot\boldsymbol{l})^2 = (l_x^2 + l_y^2 + l_z^2)^2 = l_x^4 + l_y^4 + l_z^4 + 2\gamma_1(\boldsymbol{l}), \tag{B.66}$$

the extension ratio at the $\boldsymbol{l}$-direction is given by

$$\begin{aligned}
\frac{\varepsilon(\boldsymbol{l})}{P} &= \left\{(S_{11}-S_{12})l_x^2 + S_{12}\right\} l_x^2 + \left\{(S_{11}-S_{12})l_y^2 + S_{12}\right\} l_y^2 \\
&\quad + \left\{(S_{11}-S_{12})l_z^2 + S_{12}\right\} l_z^2 + S_{44}(l_x^2 l_y^2 + l_y^2 l_z^2 + l_z^2 l_x^2) \\
&= (S_{11}-S_{12})(l_x^4 + l_y^4 + l_z^4) + S_{12}(l_x^2 + l_y^2 + l_z^2) + S_{44}(l_x^2 l_y^2 + l_y^2 l_z^2 + l_z^2 l_x^2) \\
&= (S_{11}-S_{12})(1-2\gamma_1) + S_{12} + S_{44}\gamma_1 \\
&= S_{11} + (S_{44} - 2(S_{11}-S_{12}))\gamma_1
\end{aligned} \tag{B.67}$$

or, with the stiffness constants,

$$\frac{\varepsilon(\boldsymbol{l})}{P} = \frac{C_{11}+C_{12}}{(C_{11}+2C_{12})(C_{11}-C_{12})} + \left(\frac{1}{C_{44}} - \frac{2}{C_{11}-C_{12}}\right)\gamma_1(\boldsymbol{l}) \tag{B.68}$$

With the definition

$$\gamma_2(\boldsymbol{l},\boldsymbol{m}) \equiv l_x l_y m_x m_y + l_y l_z m_y m_z + l_z l_x m_z m_x$$

and the relation

$$\begin{aligned}
0 = (\boldsymbol{l}\cdot\boldsymbol{m})^2 &= (l_x m_x + l_y m_y + l_z m_z)^2 \\
&= l_x^2 m_x^2 + l_y^2 m_y^2 + l_z^2 m_z^2 + 2\gamma_2(\boldsymbol{l},\boldsymbol{m}),
\end{aligned} \tag{B.69}$$

the extension ratio at the $\boldsymbol{m}$-direction is given by

$$\begin{aligned}
\frac{\varepsilon(\boldsymbol{m})}{P} &= \left\{(S_{11}-S_{12})l_x^2 + S_{12}\right\} m_x^2 + \left\{(S_{11}-S_{12})l_y^2 + S_{12}\right\} m_y^2 \\
&\quad + \left\{(S_{11}-S_{12})l_z^2 + S_{12}\right\} m_z^2 \\
&\quad + S_{44}(l_x l_y m_x m_y + l_y l_z m_y m_z + l_z l_x m_z m_x) \\
&= (S_{11}-S_{12})(l_x^2 m_x^2 + l_y^2 m_y^2 + l_z^2 m_z^2) + S_{12}(m_x^2 + m_y^2 + m_z^2) \\
&\quad + S_{44}(m_x^2 m_y^2 + l_y^2 l_z^2 + l_z^2 l_x^2) \\
&= (S_{11}-S_{12})(-2\gamma_2) + S_{12} + S_{44}\gamma_2 \\
&= S_{12} + \{(S_{44} - 2(S_{11}-S_{12})\}\gamma_2
\end{aligned} \tag{B.70}$$

or, with the stiffness constants,

$$\frac{\varepsilon(\boldsymbol{m})}{P} = \frac{-C_{12}}{(C_{11}+2C_{12})(C_{11}-C_{12})} + \left(\frac{1}{C_{44}} - \frac{2}{C_{11}-C_{12}}\right)\gamma_2(\boldsymbol{l},\boldsymbol{m}) \tag{B.71}$$

The Young modulus is given by

$$\begin{aligned}\frac{1}{E(\boldsymbol{l})} &\equiv \frac{\varepsilon(\boldsymbol{m})}{P} \\ &= (S_{11}-2S_{12})+(S_{44}-2(S_{11}-S_{12}))\gamma_1 \\ &= \frac{C_{11}+C_{12}}{(C_{11}+2C_{12})(C_{11}-C_{12})}+\left(\frac{1}{C_{44}}-\frac{2}{C_{11}-C_{12}}\right)\gamma_1(\boldsymbol{l}). \end{aligned} \tag{B.72}$$

The Poisson ratio is given by

$$\begin{aligned}\nu(\boldsymbol{l},\boldsymbol{m}) &\equiv -\frac{\varepsilon(\boldsymbol{m})}{\varepsilon(\boldsymbol{l})} \\ &= -\frac{S_{12}+\{(S_{44}-2(S_{11}-S_{12})\}\gamma_2}{(S_{11}-2S_{12})+\{S_{44}-2(S_{11}-S_{12})\}\gamma_1} \\ &= -\frac{\frac{-C_{12}}{(C_{11}+2C_{12})(C_{11}-C_{12})}+\left(\frac{1}{C_{44}}-\frac{2}{C_{11}-C_{12}}\right)\gamma_2(\boldsymbol{l},\boldsymbol{m})}{\frac{C_{11}+C_{12}}{(C_{11}+2C_{12})(C_{11}-C_{12})}+\left(\frac{1}{C_{44}}-\frac{2}{C_{11}-C_{12}}\right)\gamma_1(\boldsymbol{l})} \end{aligned} \tag{B.73}$$

In the case with $\boldsymbol{l}=(1,0,0)$ and $\boldsymbol{m}=(0,1,0)$, the values $\gamma_1=0,\gamma_2=0$ are obtained and Eqs. (B.72),(B.73) are reduced to Eqs. (B.39), (B.40). From Eqs. (B.72),(B.73), the isotropic elastic properties will be given by

$$C_{44}=\frac{C_{11}-C_{12}}{2}. \tag{B.74}$$

The anisotropic Young modulus in silicon is shown in Table B.1.

	C_{11}	C_{12}	C_{44}	E_{100}	E_{110}	E_{111}
Si	166	64	80	130	170	189

Table B.1: The elastic constants C_{11}, C_{12}, C_{44} and the anisotropic Young moduli $E_{100}, E_{110}, E_{111}$ in silicon are shown in the unit of GPa. The anisotropic Young moduli are calculated by Eq. (B.72). Note that the (100),(110) and (111) directions correspond to $\gamma_1=0$,$\gamma_1=1/4$ and $\gamma_1=1/3$, respectively.

B.2 Simple classical model in tetrahedral structure

Here we demonstrate how the anisotropic elastic property in the tetrahedral structure is derived from a microscopic model. We consider a simple classical model, which is consist of the two harmonic potentials in the bond stretching mode and the bond bending mode [62]. The model contains the two parameters and we will fit the two parameters so as to reproduce the bulk modulus B and the shear modulus $C_{\rm s}(\equiv C_{11}-C_{12})$. Then the shear modulus C_{44} will be determined from the present model and compared to the experimental value.

Figure B.2 shows the tetrahedral structure with the two shear modes. The atom placed at (0,0,0) and the four nearest neighbor atoms are picked out

$$\begin{aligned}
&\text{O} \quad \frac{l_0}{\sqrt{3}}(0,0,0)\\
&\text{A} \quad \frac{l_0}{\sqrt{3}}(1,1,1)\\
&\text{B} \quad \frac{l_0}{\sqrt{3}}(1,-1,-1)\\
&\text{C} \quad \frac{l_0}{\sqrt{3}}(-1,-1,1)\\
&\text{D} \quad \frac{l_0}{\sqrt{3}}(-1,1,-1).
\end{aligned}$$

We will consider the above five atoms that includes the four bonds (OA,OB,OC,OD) and six bond angles (AOB,BOC,COD,DOA). The total strain energy per atom is given by

$$U=\frac{1}{2}\sum_{i:\text{bond}}^{4}\frac{\alpha}{2}\left(\frac{\delta l_i}{l_0}\right)^2+\sum_{j:\text{bond angle}}^{6}\frac{\beta}{2}\left(\delta\theta_j\right)^2. \tag{B.75}$$

The factor 1/2 in the first term cancels the double counting of the bond stretching energy; For example, the stretching energy of the bond OA is counted as the energy per atom for the atom O and the atom A.

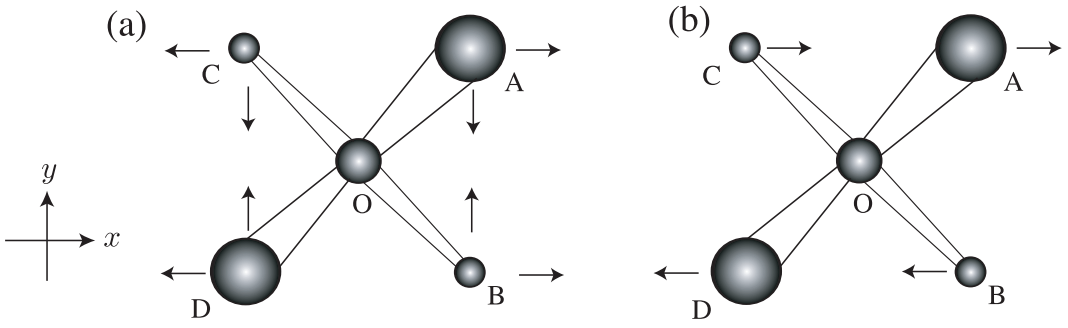


Figure B.2: Top view of the tetrahedral structure. The arrows indicate the x and y components of the deformation in (a) the shear mode of $C_{11}-C_{12}$ or (b) the shear mode of C_{44}. See the text for the coordinates of each atom.

Bulk modulus

The bulk modulus is derived, when the deformed atomic coordinates are chosen as

$$\begin{aligned}
\mathrm{O} \quad & \frac{l_0}{\sqrt{3}}(0,0,0) \\
\mathrm{A} \quad & (1+\varepsilon)\frac{l_0}{\sqrt{3}}(1,1,1) \\
\mathrm{B} \quad & (1+\varepsilon)\frac{l_0}{\sqrt{3}}(1,-1,-1) \\
\mathrm{C} \quad & (1+\varepsilon)\frac{l_0}{\sqrt{3}}(-1,-1,1) \\
\mathrm{D} \quad & (1+\varepsilon)\frac{l_0}{\sqrt{3}}(-1,1,-1).
\end{aligned}$$

The corresponding strain energy is

$$U = U_1 \equiv \frac{B}{2}(3\varepsilon)^2 v_0 = \frac{9Bv_0}{2}\varepsilon^2 \tag{B.76}$$

per atom, due to the definition of the bulk modulus. In Eq. (B.75), on the other hand, the four bond length is stretched by $\delta l_i/l_0 = \varepsilon$ and no bond angle is changed. Thus the strain energy U_1 should be written as

$$U_1 = \frac{1}{2} \times 4 \times \frac{\alpha}{2}\varepsilon^2 = \varepsilon^2\alpha. \tag{B.77}$$

Now we obtain the relation between the microscopic parameter α and the Bulk modulus B as

$$\alpha = \frac{9Bv_0}{2} \tag{B.78}$$

Shear modulus $C_{11} - C_{12}$

The shear modulus $C_\mathrm{s} \equiv C_{11} - C_{12}$ is derived, when the deformed atomic coordinates are chosen as

$$\begin{aligned}
\mathrm{O} \quad & \frac{l_0}{\sqrt{3}}(0,0,0) \\
\mathrm{A} \quad & \frac{l_0}{\sqrt{3}}(+1+\varepsilon,+1-\varepsilon,+1) \\
\mathrm{B} \quad & \frac{l_0}{\sqrt{3}}(+1+\varepsilon,-1-\varepsilon,-1) \\
\mathrm{C} \quad & \frac{l_0}{\sqrt{3}}(-1-\varepsilon,-1+\varepsilon,+1) \\
\mathrm{D} \quad & \frac{l_0}{\sqrt{3}}(-1-\varepsilon,+1-\varepsilon,-1).
\end{aligned}$$

Within the linear order of ε, no bond length is changed. The change of the six bond angles are classified into three types;

$$\begin{aligned} & \cos(\mathrm{AOC}) = \cos(\mathrm{BOD}) \\ = \; & \frac{1}{3}\left\{(1+\varepsilon)(-1-\varepsilon)+(1-\varepsilon)(-1+\varepsilon)+1\right\} \\ = \; & -\frac{1}{3}+O\left(\varepsilon^2\right) \end{aligned} \tag{B.79}$$

$$\begin{aligned} & \cos(\mathrm{AOB}) = \cos(\mathrm{COD}) \\ = \; & \frac{1}{3}\left\{(1+\varepsilon)(1+\varepsilon)+(1-\varepsilon)(-1+\varepsilon)-1\right\} \\ = \; & -\frac{1}{3}\left(1-4\varepsilon\right)+O\left(\varepsilon^2\right) \end{aligned} \tag{B.80}$$

$$\begin{aligned} & \cos(\mathrm{BOC}) = \cos(\mathrm{AOD}) \\ = \; & \frac{1}{3}\left\{(1+\varepsilon)(-1-\varepsilon)+(-1+\varepsilon)(-1+\varepsilon)-1\right\} \\ = \; & -\frac{1}{3}\left(1+4\varepsilon\right)+O\left(\varepsilon^2\right) \end{aligned} \tag{B.81}$$

We recall that the bond angle in the ideal crystal θ_0 gives

$$\cos\theta_0 = -\frac{1}{3}, \quad \sin\theta_0 = \frac{2\sqrt{2}}{3} \tag{B.82}$$

and $\theta_0 \approx 109.47$. For a small deviation $\delta\theta$, we obtain

$$\begin{aligned} \cos(\theta_0+\delta\theta) &= \cos\theta_0\cos(\delta\theta)-\sin\theta_0\sin(\delta\theta) \\ &\approx -\frac{1}{3}-\frac{2\sqrt{2}}{3}\delta\theta \\ &= -\frac{1}{3}\left(1+2\sqrt{2}\delta\theta\right). \end{aligned} \tag{B.83}$$

When the above relation is compared with Eq. (B.80) or (B.81), the four bond angles are changed by the amplitude of

$$|\delta\theta| = \sqrt{2}\varepsilon \tag{B.84}$$

The corresponding strain energy per atom is given by

$$U = U_2 \equiv 4 \times \frac{\beta}{2}(\sqrt{2}\varepsilon)^2 = 4\beta\varepsilon^2, \tag{B.85}$$

which is compared with

$$U_2 \equiv C_{\mathrm{s}}\varepsilon^2 v_0, \tag{B.86}$$

Therefore we obtain the relation between the microscopic parameter β and the elastic constant C_{s}

$$\beta = \frac{C_{\mathrm{s}}}{4}v_0 \tag{B.87}$$

Shear modulus C_{44}

The shear modulus C_{44}, when the deformed atomic coordinates are chosen as

$$\begin{aligned}
\mathrm{O} \quad & \frac{l_0}{\sqrt{3}}(0,0,\xi\varepsilon) \\
\mathrm{A} \quad & \frac{l_0}{\sqrt{3}}(+1+\varepsilon,+1,+1) \\
\mathrm{B} \quad & \frac{l_0}{\sqrt{3}}(+1-\varepsilon,-1,-1) \\
\mathrm{C} \quad & \frac{l_0}{\sqrt{3}}(-1-\varepsilon,-1,+1) \\
\mathrm{D} \quad & \frac{l_0}{\sqrt{3}}(-1+\varepsilon,+1,-1).
\end{aligned}$$

Here we introduce an internal strain parameter ξ [133, 81, 62]. We will find that the parameter is positive. The corresponding strain energy per atom is given by

$$U_3 \equiv \frac{C_{44}}{2}\varepsilon^2 v_0. \tag{B.88}$$

The four bond lengths are classified into two types;

$$\begin{aligned}
& (\mathrm{OA}) = (\mathrm{OC}) \\
= \; & \frac{l_0}{\sqrt{3}}\left\{(1+\varepsilon)^2+1+(1-\xi\varepsilon)^2\right\}^{\frac{1}{2}} \\
\approx \; & \frac{l_0}{\sqrt{3}}\left\{1+2\varepsilon+1+1-2\xi\varepsilon\right\}^{\frac{1}{2}} \\
= \; & l_0\left\{1+\frac{2(1-\xi)}{3}\varepsilon\right\}^{\frac{1}{2}} \\
\approx \; & l_0\left\{1+\frac{(1-\xi)}{3}\varepsilon\right\} \qquad \text{(B.89)} \\
& (\mathrm{OB}) = (\mathrm{OD}) \\
= \; & \frac{l_0}{\sqrt{3}}\left\{(1-\varepsilon)^2+1+(1+\xi\varepsilon)^2\right\}^{\frac{1}{2}} \\
\approx \; & l_0\left\{1-\frac{(1-\xi)}{3}\varepsilon\right\}. \qquad \text{(B.90)}
\end{aligned}$$

From the above relations, the bond stretching energy is contributed by four bonds and is given by

$$\frac{1}{2}\sum_{i:\mathrm{bond}}^{4}\frac{\alpha}{2}\left(\frac{\delta l_i}{l_0}\right)^2 \Rightarrow \frac{1}{2}\times 4\times\frac{\alpha}{2}\times\left\{\frac{1-\xi}{3}\varepsilon\right\}^2 = \frac{\alpha(1-\xi)^2}{9}\varepsilon^2. \tag{B.91}$$

On the other hand, the bond angles are classified into three types; The first type is the angle between two shorter bonds, that is, the angle (BOD). We calculate the

inner product

$$\begin{aligned}\left(\overrightarrow{\mathrm{OB}}\cdot\overrightarrow{\mathrm{OD}}\right) &= \frac{l_0^2}{3}\left\{(1-\varepsilon)(-1+\varepsilon)+(-1)(+1)+(-1-\xi\varepsilon)(-1-\xi\varepsilon)\right\}\\ &\approx -\frac{l_0^2}{3}\left\{1-2(1+\xi)\varepsilon\right\} \qquad \text{(B.92)}\end{aligned}$$

and the product of the bond lengths,using $(B.90)$,

$$|\mathrm{OB}|\,|\mathrm{OD}| = l_0^2\left\{1-\frac{(1-\xi)}{3}\varepsilon\right\}^2 \approx l_0^2\left\{1-\frac{2(1-\xi)}{3}\varepsilon\right\}. \qquad \text{(B.93)}$$

From the above relations, we obtain

$$\begin{aligned}\cos(\mathrm{BOD}) &= \frac{\overrightarrow{\mathrm{OB}}\cdot\overrightarrow{\mathrm{OD}}}{|\mathrm{OB}|\,|\mathrm{OD}|}\\ &= -\frac{1}{3}\frac{1-2(1+\xi)\varepsilon}{1-\frac{2(1-\xi)}{3}\varepsilon}\\ &= -\frac{1}{3}\left\{1-2(1+\xi)\varepsilon\right\}\left\{1+\frac{2(1-\xi)}{3}\varepsilon\right\}\\ &= -\frac{1}{3}\left\{1-\frac{4}{3}(1+2\xi)\varepsilon\right\}. \qquad \text{(B.94)}\end{aligned}$$

The second type is the angle between two longer bonds, that is, the angle (AOC). We calculate the inner product

$$\begin{aligned}&\left(\overrightarrow{\mathrm{OA}}\cdot\overrightarrow{\mathrm{OC}}\right)\\ &= \frac{l_0^2}{3}\left\{(1+\varepsilon)(-1-\varepsilon)+(+1)(-1)+(1-\xi\varepsilon)(1-\xi\varepsilon)\right\}\\ &\approx -\frac{l_0^2}{3}\left\{1+2(1+\xi)\varepsilon\right\} \qquad \text{(B.95)}\end{aligned}$$

and the product of the bond lengths,using $(B.89)$,

$$|\mathrm{OA}|\,|\mathrm{OC}| = l_0^2\left\{1-\frac{(1+\xi)}{3}\varepsilon\right\}^2 \approx l_0^2\left\{1+\frac{2(1-\xi)}{3}\varepsilon\right\}. \qquad \text{(B.96)}$$

From the above relations, we obtain

$$\begin{aligned}\cos(\mathrm{AOC}) &= \frac{\overrightarrow{\mathrm{OA}}\cdot\overrightarrow{\mathrm{OC}}}{|\mathrm{OA}|\,|\mathrm{OC}|}\\ &= -\frac{1}{3}\frac{1+2(1+\xi)\varepsilon}{1+\frac{2(1-\xi)}{3}\varepsilon}\\ &= -\frac{1}{3}\left\{1+2(1+\xi)\varepsilon\right\}\left\{1-\frac{2(1-\xi)}{3}\varepsilon\right\}\\ &= -\frac{1}{3}\left\{1+\frac{4}{3}(1+2\xi)\varepsilon\right\}. \qquad \text{(B.97)}\end{aligned}$$

The other four bond angles, (AOB),(DOA),(COD),(BOC), are classified by the bond angle between a shorter bond and a longer bond. The above angles are not changed within the linear order, which is proved, as follows. For example, we calculate cos(AOB) using the inner product

$$\begin{aligned} & \left(\overrightarrow{\mathrm{OA}} \cdot \overrightarrow{\mathrm{OB}}\right) \\ = & \frac{1}{3}\{(1+\varepsilon)(1-\varepsilon)+(+1)(-1)+(1-\xi\varepsilon)(-1-\xi\varepsilon)\} \\ = & -\frac{1}{\sqrt{3}}\left\{1+O(\varepsilon^2)\right\} \end{aligned} \tag{B.98}$$

and the product of the bond lengths

$$\begin{aligned} & |\mathrm{OA}|\,|\mathrm{OB}| \\ = & l_0^2\left\{1-\frac{(1+\xi)}{3}\varepsilon\right\}\times\left\{1+\frac{(1+\xi)}{3}\varepsilon\right\} \approx l_0^2\left\{1+O(\varepsilon^2)\right\}. \end{aligned} \tag{B.99}$$

The above relations result in

$$\cos(\mathrm{AOB}) = \frac{\overrightarrow{\mathrm{OA}} \cdot \overrightarrow{\mathrm{OB}}}{|\mathrm{OA}|\,|\mathrm{OB}|} = -\frac{1}{3} + O\left(\varepsilon^2\right). \tag{B.100}$$

From the above calculations of the six bond angles, we can see that the bond bending energy is contributed by *two* bond angles whose amplitudes are the same value of

$$|\delta\theta| = \frac{\sqrt{2}}{3}(1+2\xi)\varepsilon, \tag{B.101}$$

where we use Eq. (B.83). The corresponding bond bending energy is given by

$$\sum_{j:\text{bond angle}}^{6} \frac{\beta}{2}\left(\delta\theta_j\right)^2 \Rightarrow 2\times\frac{\beta}{2}\left(\frac{\sqrt{2}}{3}(1+2\xi)\varepsilon\right)^2 = \frac{2\beta}{9}(1+2\xi)^2\varepsilon^2. \tag{B.102}$$

The strain energy due to C_{44} is the sum of Eq. (B.91) and Eq. (B.102), which is given, as a function of ξ by

$$U_3(\xi) = \frac{\alpha(1-\xi)^2}{9}\varepsilon^2. + \frac{2\beta}{9}(1+2\xi)^2\varepsilon^2. \tag{B.103}$$

The equilibrium strain energy is determined by relaxing the internal strain parameter ξ;

$$U_3'(\xi) = 0. \tag{B.104}$$

The relaxed value is denoted as ζ and is given by

$$\begin{aligned} 0 &= \frac{9}{\varepsilon^2}U_3'(\zeta) \\ &= 2\alpha(1-\zeta)(-1)+4\beta(1+2\zeta)2 \\ &= -2\{\alpha(1-\zeta)-4\beta(1+2\zeta)\} \\ &= -2\{(\alpha-4\beta)-\zeta(\alpha+8\beta)\}. \end{aligned} \tag{B.105}$$

As results, we obtain

$$\zeta = \frac{\alpha - 4\beta}{\alpha + 8\beta}. \tag{B.106}$$

Now we would like to calculate the relaxed strain energy $U_3(\zeta)$. As preparations, we calculate

$$1 - \zeta = \frac{\alpha + 8\beta}{\alpha + 8\beta} - \frac{\alpha - 4\beta}{\alpha + 8\beta} = \frac{12\beta}{\alpha + 8\beta} \tag{B.107}$$

$$1 + 2\zeta = \frac{\alpha + 8\beta}{\alpha + 8\beta} + \frac{2\alpha - 8\beta}{\alpha + 8\beta} = \frac{3\alpha}{\alpha + 8\beta}. \tag{B.108}$$

When the above relations are introduced into Eq. (B.103),

$$\begin{aligned} U_3(\zeta) &= \frac{\alpha(1-\zeta)^2}{9}\varepsilon^2. + \frac{2\beta}{9}(1+2\zeta)^2\varepsilon^2 \\ &= \frac{\varepsilon^2}{9}\left\{\alpha\left(\frac{12\beta}{\alpha+8\beta}\right)^2 + 2\beta\left(\frac{3\alpha}{\alpha+8\beta}\right)^2\right\} \\ &= \frac{\varepsilon^2}{9(\alpha+8\beta)^2}\left\{144\alpha\beta^2 + 18\alpha^2\beta\right\} \\ &= \frac{\varepsilon^2}{9(\alpha+8\beta)^2} \times 18\alpha\beta(8\beta+\alpha) \\ &= \frac{2\alpha\beta\varepsilon^2}{\alpha+8\beta}. \end{aligned} \tag{B.109}$$

We rewrite Eq. (B.78) and Eq. (B.87)

$$\alpha = \frac{9Bv_0}{2}, \quad \beta = \frac{C_{\rm s}}{4}v_0. \tag{B.110}$$

From the above relation, the internal strain and the relaxed strain energy is given by

$$\zeta = \frac{\alpha - 4\beta}{\alpha + 8\beta} = \frac{9B - 2C_{\rm s}}{9B + 4C_{\rm s}}, \tag{B.111}$$

and

$$U_3(\zeta) = \frac{2\alpha\beta\varepsilon^2}{\alpha+8\beta} = \frac{2\frac{9B}{2}\frac{C_{\rm s}}{4}}{\frac{9B}{2}+8\frac{C_{\rm s}}{4}}\varepsilon^2 = \frac{1}{2}\frac{9BC_{\rm s}}{9B+4C_{\rm s}}\varepsilon^2. \tag{B.112}$$

Since the above energy should be equal to Eq. (B.88), we obtain an important relation

$$C_{44} = \frac{9BC_{\rm s}}{9B + 4C_{\rm s}}. \tag{B.113}$$

Moreover, an theoretical quantity $C_{44}^{(0)}$ is often defined as the shear constant for the strain *without* the internal strain ($\xi = 0$), which results, using Eq. (B.103), in

$$\frac{C_{44}^{(0)}}{2}\varepsilon^2 v_0 \equiv U_3(\xi{=}0) = \frac{\alpha}{9}\varepsilon^2 + \frac{2\beta}{9}\varepsilon^2. \tag{B.114}$$

Using the above relation and Eq. (B.110),

$$C_{44}^{(0)} = \frac{1}{9v_0}(2\alpha + 4\beta) = \frac{1}{9}\left(2\frac{9B}{2} + 4\frac{C_{\rm s}}{4}\right) = B + \frac{1}{9}C_{\rm s}. \tag{B.115}$$

Equations (B.111), (B.113), (B.115) are the resultant relations.

Discussion

The above two-parameter classical model is applied to silicon. In silicon case, the bulk modulus nearly equals to the shear modulus

$$B \approx C_{\rm s} \approx 100\,[\text{GPa}], \tag{B.116}$$

which is adopted as the input values. In the case with $B \to C_{\rm s}$, Eqs. (B.111), (B.113), (B.115) are reduced to

$$\zeta \;\to\; \frac{7}{13} \approx 0.54 \tag{B.117}$$

$$\frac{C_{44}}{C_{\rm s}} \;\to\; \frac{9}{13} \approx 0.69 \tag{B.118}$$

$$\frac{C_{44}^{(0)}}{C_{\rm s}} \;\to\; \frac{10}{9} \approx 1.11 \tag{B.119}$$

The results are shown in Table B.2 with *ab initio* and experimental values. The present simple model reproduces the experimental values well, which implies that the elastic constants can be understood by the present simple model, that is, the model with the harmonic potentials of the bond stretching and bond bending.

Here we add several comments; (a) Keating gives the following expressions [134]:

$$\zeta = \frac{2C_{12}}{C_{11}+C_{12}} = \frac{3B - C_{\rm s}}{3B + C_{\rm s}} \tag{B.120}$$

$$C_{44} = \frac{(C_{11}-C_{12})(C_{11}+3C_{12})}{2(C_{11}+C_{12})} = \frac{3BC_{\rm s}}{3B+C_{\rm s}}, \tag{B.121}$$

which are similar to Eqs. (B.111) and (B.113), respectively. (b) The internal strain parameter ζ is determined by the competition of the bond stretching energy and the bond bending energy. In limiting cases, the value will be given as

$$\zeta \;\to\; 1, \quad (\alpha \gg \beta \text{ or } B \gg C_{\rm s}) \tag{B.122}$$

$$\zeta \;\to\; -\frac{1}{2}, \quad (\alpha \ll \beta \text{ or } B \ll C_{\rm s}). \tag{B.123}$$

In silicon case, the value of ζ is positive, which means that the bond stretching energy is more important than the bond bending energy. In other words, the bond is harder to stretch but easier to bend. (c) Since the present simple model is a three-body interatomic potential, the model does not reproduce the energy difference of the diamond structure and the wurzite structure. So as to reproduce the above difference, the energy term for the dihedral angles should be considered.

	Simple model	*ab initio*[81]	Exp.
B [GPa]	100 (input)	92.0	97.8
$C_{\rm s}$[GPa]	100 (input)	98.0	101.2
C_{44} [GPa]	69	85.0	79.6
$C_{44}^{(0)}$ [GPa]	111	111	-
ζ	0.54	0.53	0.54 [135]

Table B.2: The elastic constants and the internal strain parameter ζ in silicon, where $B \equiv (C_{11}+2C_{12})/3, C_{\rm s} \equiv C_{11}-C_{12}$. In the 'simple model', $B = C_{\rm s} = 100$ GPa are used as the input values and the other quantities are given by Eqs. (B.111), (B.113), (B.115).

B.3 Theory in isotropic medium

This appendix is devoted to the theory of elasticity in isotropic (3D) medium. See Refs. [132, 136] as standard textbooks. As explained in Section B.1, the number of the independent elastic parameters is reduced to two, in isotropic media. We redefine the strain energy with two parameters as

$$\begin{aligned} U &\equiv U_0 + \frac{\lambda}{2}\left(\sum_i u_{ii}\right)^2 + \mu \sum_{ij} u_{ij}^2 \\ &= U_0 + \frac{\lambda}{2}\left(\mathrm{Tr}[u]\right)^2 + \mu \sum_{ij} u_{ij}^2. \end{aligned} \tag{B.124}$$

The parameters λ and μ are called Lame parameters. Here the trace

$$\mathrm{Tr}[u] = u_{xx} + u_{yy} + u_{zz} = \frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z} = \mathrm{div}\boldsymbol{u} \tag{B.125}$$

corresponds to the ratio of the volume expansion. Note that, there must be some conditions on the components of the strain tensor, because they are defined as partial derivatives of the components of the displacement vector $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{r})$. These conditions are called the condition of compatibility [132, 136].

Stress and strain

The stress tensor σ is given by

$$\sigma_{ij} \equiv \left(\frac{\partial U}{\partial u_{ij}}\right)_{9\,\text{variables}} = \lambda\left(\mathrm{Tr}[u]\right)\delta_{ij} + 2\mu u_{ij}. \tag{B.126}$$

The trace of Eq. (B.126) is reduced to

$$\mathrm{Tr}[\sigma] = \lambda\left(\mathrm{Tr}[u]\right)\sum_{ij}\delta_{ij} + 2\mu\mathrm{Tr}[u] = (3\lambda + 2\mu)\mathrm{Tr}[u]. \tag{B.127}$$

Using Eq. (B.127), Eq. (B.126) is solved for u_{ij} as

$$u_{ij} = \frac{1}{2\mu}\left[\sigma_{ij} - \frac{\lambda\mathrm{Tr}[\sigma]}{3\lambda + 2\mu}\delta_{ij}\right]. \tag{B.128}$$

Bulk and shear modulus

If we define the shear stain matrix as

$$\tilde{u}_{ij} \equiv u_{ij} - \frac{\mathrm{Tr}[u]}{3}\delta_{ij}, \tag{B.129}$$

its trace is equal to zero

$$\mathrm{Tr}[\tilde{u}] = 0. \tag{B.130}$$

Using the above relations, one obtains

$$\begin{aligned}\sum_{ij} u_{ij}^2 &= \sum_{ij}\left(\tilde{u}_{ij}+\frac{\mathrm{Tr}[u]}{3}\delta_{ij}\right)^2 \\ &= \sum_{ij}(\tilde{u}_{ij})^2+\frac{2\mathrm{Tr}[u]}{3}\sum_{ij}\delta_{ij}\tilde{u}_{ij}+\left(\frac{\mathrm{Tr}[u]}{3}\right)^2\sum_{ij}\delta_{ij}^2 \\ &= \sum_{ij}(\tilde{u}_{ij})^2+0+\left(\frac{\mathrm{Tr}[u]}{3}\right)^2 3 \\ &= \sum_{ij}(\tilde{u}_{ij})^2+\frac{(\mathrm{Tr}[u])^2}{3}. \end{aligned} \tag{B.131}$$

From above all, the deformation energy is written as

$$U = U_0+\frac{1}{2}\left(\lambda+\frac{2}{3}\mu\right)(\mathrm{Tr}[u])^2+\mu\sum_{ij}\tilde{u}_{ij}^2. \tag{B.132}$$

Here the first term corresponds to the volume expansions and the second term corresponds to the shear strains. For the energy stability, the coefficients must satisfy

$$B \equiv \lambda+\frac{2}{3}\mu > 0, \quad \mu > 0. \tag{B.133}$$

Here B or μ is called the bulk or shear modulus, respectively.

Young modulus and Poisson ratio

If the components of stress are chosen to be zero except σ_{xx}, Eq. (B.128) gives the non-zero components of strain as

$$u_{xx} = \frac{\lambda+\mu}{\mu(3\lambda+2\mu)}\sigma_{xx} \tag{B.134}$$

$$u_{yy}=u_{zz} = -\frac{\lambda}{2\mu(3\lambda+2\mu)}\sigma_{xx}. \tag{B.135}$$

The Young modulus E and the Poisson ratio ν are given as

$$E \equiv \frac{\sigma_{xx}}{u_{xx}} = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu} \tag{B.136}$$

$$\nu \equiv -\frac{u_{yy}}{u_{xx}} = \frac{\lambda}{2(\lambda+\mu)}. \tag{B.137}$$

Since the parameter set (λ,μ) can be written by the parameter set (E,ν)

$$\begin{aligned}\lambda &= \frac{\nu E}{(1-2\nu)(\nu+1)} \\ \mu &= \frac{E}{2(\nu+1)}, \end{aligned} \tag{B.138}$$

Eq. (B.126) gives

$$\sigma_{xx} = \frac{E}{(1+\mu)}\left\{\frac{1-\nu}{1-2\nu}u_{xx} + \frac{\nu}{1-2\nu}(u_{yy}+u_{zz})\right\} \tag{B.139}$$

$$\sigma_{xy} = \frac{E}{(1+\mu)}u_{xy}. \tag{B.140}$$

When Eqs. (B.139), (B.140) are compared with Eqs. (B.12), (B.13), we obtain

$$C_{11} = \frac{E}{(1+\mu)}\frac{1-\nu}{1-2\nu} \tag{B.141}$$

$$C_{12} = \frac{E}{(1+\mu)}\frac{\nu}{1-2\nu} \tag{B.142}$$

$$C_{44} = \frac{C_{11}-C_{12}}{2} = \frac{E}{2(1+\nu)}. \tag{B.143}$$

From Eq. (B.137), the Poisson ratio is rewritten as

$$\nu = \frac{1}{2}\frac{\lambda}{\lambda+\mu} = \frac{1}{2}\frac{B-\frac{2}{3}\mu}{B+\frac{1}{3}\mu} = \frac{1}{2}\frac{1-\frac{2}{3}\frac{\mu}{B}}{1+\frac{1}{3}\frac{\mu}{B}}. \tag{B.144}$$

Since the parameters μ and B are positive, the Poisson ratio ν must satisfy

$$-1 \leq \nu \leq \frac{1}{2}. \tag{B.145}$$

In most experiments, the Poisson ratio is positive ($0 < \nu < 1/2$).

Equation of motion

The equation of motion for a local region is given by

$$\rho\ddot{u}_i = \sum_j \frac{\partial \sigma_{ij}}{\partial x_j}, \tag{B.146}$$

where the constant ρ is the density. If the right hand side is denoted as f_i, the vector (f_x, f_y, f_z) is the force imposed on the local region. From Eq. (B.126), the right hand side of (B.146) is given as

$$\begin{aligned}
\sum_j \frac{\partial \sigma_{ij}}{\partial x_j} &= \lambda\frac{\partial}{\partial x_i}(\mathrm{Tr}[u]) + 2\mu\sum_j \frac{\partial u_{ij}}{\partial x_j} \\
&= \lambda\frac{\partial}{\partial x_i}(\mathrm{Tr}[u]) + \mu\sum_j \frac{\partial}{\partial x_j}\left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right) \\
&= \lambda\frac{\partial}{\partial x_i}(\mathrm{Tr}[u]) + \mu\frac{\partial}{\partial x_i}\sum_j \frac{\partial u_j}{\partial x_j} + \mu\sum_j \frac{\partial^2 u_i}{\partial x_j^2} \\
&= (\lambda+\mu)\frac{\partial}{\partial x_i}(\mathrm{Tr}[u]) + \mu\Delta u_i.
\end{aligned} \tag{B.147}$$

Therefore, Eq. (B.146) is given, in a vector form, as

$$\rho\ddot{\boldsymbol{u}} = (\lambda+\mu)\mathrm{grad}\,\mathrm{div}\boldsymbol{u} + \mu\Delta\boldsymbol{u}. \tag{B.148}$$

Biharmonic property of the displacement vector

From Eq. (B.148), the balance equation is given as

$$0 = (\lambda + \mu)\mathrm{grad}\,\mathrm{div}\boldsymbol{u} + \mu\Delta\boldsymbol{u} \tag{B.149}$$

or, with the relation $\Delta = \mathrm{grad}\,\mathrm{div} - \mathrm{rot}\,\mathrm{rot}$,

$$0 = (\lambda + 2\mu)\mathrm{grad}\,\mathrm{div}\boldsymbol{u} - \mu\,\mathrm{rot}\,\mathrm{rot}\boldsymbol{u}. \tag{B.150}$$

When we take the divergence of Eq. (B.150) and use the relation $\mathrm{div}\,\mathrm{rot} = 0$, we find that $\mathrm{div}\boldsymbol{u}$ is harmonic;

$$0 = \mathrm{div}\,\mathrm{grad}\,\mathrm{div}\boldsymbol{u} = \Delta\mathrm{div}\boldsymbol{u}. \tag{B.151}$$

If the Laplacian Δ is operated on Eq. (B.149), we obtain

$$\begin{aligned} 0 &= (\lambda + \mu)\Delta\mathrm{grad}\,\mathrm{div}\boldsymbol{u} + \mu\Delta^2\boldsymbol{u} \\ &= (\lambda + \mu)\mathrm{grad}\,\Delta\mathrm{div}\boldsymbol{u} + \mu\Delta^2\boldsymbol{u}, \end{aligned} \tag{B.152}$$

where the second equality is given by the relation $\Delta\,\mathrm{grad} = \mathrm{grad}\,\Delta$. From Eq. (B.151) and Eq. (B.152), we conclude that $\boldsymbol{u}$ is biharmonic;

$$\Delta^2\boldsymbol{u} = 0. \tag{B.153}$$

Elastic wave

Now two velocity parameters (c_l, c_t) are introduced by the definitions

$$c_l = \sqrt{\frac{\lambda + 2\mu}{\rho}} = \sqrt{\frac{3B + 4\mu}{3\rho}} \tag{B.154}$$

$$c_t = \sqrt{\frac{\mu}{\rho}}. \tag{B.155}$$

Here the inside of the square root is positive, because of $B, \mu > 0$. The parameter set (c_l, c_t) can be written by the parameter set (E, ν) as

$$c_l = \sqrt{\frac{E(1-\nu)}{\rho(1+\nu)(1-2\nu)}} \tag{B.156}$$

$$c_t = \sqrt{\frac{E}{2\rho(1+\nu)}}. \tag{B.157}$$

Now the ratio between the two velocity parameters are determined only by the Poisson ratio ν as

$$\frac{c_t}{c_l} = \sqrt{\frac{1-2\nu}{2(1-\nu)}}. \tag{B.158}$$

Note that $c_l > \sqrt{2}c_t$ within the range of $0 < \nu < 1/2$. We also find that

$$\lambda = \rho\left(c_l^2 - 2c_t^2\right) \tag{B.159}$$
$$\mu = \rho c_t^2 \tag{B.160}$$
$$E = \rho c_t^2 \frac{3c_l^2 - 4c_t^2}{c_l^2 - c_t^2} \tag{B.161}$$
$$\nu = \frac{c_l^2 - 2c_t^2}{2(c_l^2 - c_t^2)}. \tag{B.162}$$

Several useful relations are added

$$\lambda + 2\mu = \rho c_l^2 \tag{B.163}$$
$$1 - \nu = \frac{c_l^2}{2(c_l^2 - c_t^2)}. \tag{B.164}$$

The equation of motion of Eq. (B.148) is written, with c_l, c_t, as

$$\ddot{\boldsymbol{u}} = c_t^2 \Delta \boldsymbol{u} + (c_l^2 - c_t^2)\mathrm{grad}\,\mathrm{div}\boldsymbol{u}. \tag{B.165}$$

From vector analysis, a vector field can be decomposed into two vector fields

$$\boldsymbol{u} = \boldsymbol{u}_\mathrm{l} + \boldsymbol{u}_\mathrm{t} \tag{B.166}$$

where

$$\mathrm{div}\boldsymbol{u}_\mathrm{t} = 0 \tag{B.167}$$
$$\mathrm{rot}\boldsymbol{u}_\mathrm{l} = 0. \tag{B.168}$$

When Eq. (B.166) is substituted to Eq. (B.165),

$$0 = \ddot{\boldsymbol{u}}_\mathrm{l} + \ddot{\boldsymbol{u}}_\mathrm{t} - c_t^2 \Delta \boldsymbol{u}_\mathrm{l} - c_t^2 \Delta \boldsymbol{u}_\mathrm{t} - (c_l^2 - c_t^2)\mathrm{grad}\,\mathrm{div}\boldsymbol{u}_\mathrm{l}. \tag{B.169}$$

If we take the divergence of Eq. (B.169) and use the relation $\Delta\mathrm{div} = \mathrm{div}\Delta$, we obtain

$$\begin{aligned} 0 &= \mathrm{div}\,\ddot{\boldsymbol{u}}_\mathrm{l} - c_t^2 \Delta \mathrm{div}\,\boldsymbol{u}_\mathrm{l} - (c_l^2 - c_t^2)\Delta\,\mathrm{div}\boldsymbol{u}_\mathrm{l} \\ &= \mathrm{div}\left(\ddot{\boldsymbol{u}}_\mathrm{l} - c_l^2 \boldsymbol{u}_\mathrm{l}\right). \end{aligned} \tag{B.170}$$

On the other hand, Eq. (B.168) gives relation

$$0 = \mathrm{rot}\left(\ddot{\boldsymbol{u}}_\mathrm{l} - c_l^2 \Delta \boldsymbol{u}_\mathrm{l}\right). \tag{B.171}$$

In general, if a vector $\boldsymbol{A}$ satisfies $\mathrm{div}\boldsymbol{A} = 0$ and $\mathrm{rot}\boldsymbol{A} = 0$, the vector $\boldsymbol{A}$ should be a constant vector. Since the elastic wave has the trivial solution of $\boldsymbol{u}_\mathrm{l} = 0$, Eqs. (B.170),(B.171) give

$$\ddot{\boldsymbol{u}}_\mathrm{l} - c_l^2 \Delta \boldsymbol{u}_\mathrm{l} = 0. \tag{B.172}$$

If we take the rotation of (B.169) and use the relations $\Delta\mathrm{rot} = \mathrm{rot}\Delta$ and $\mathrm{rot\,grad} = 0$, we obtain

$$\begin{aligned} 0 &= \mathrm{rot}\,\ddot{\boldsymbol{u}}_{\mathrm{t}} - c_t^2 \Delta \mathrm{rot}\,\boldsymbol{u}_{\mathrm{l}} \\ &= \mathrm{rot}\left(\ddot{\boldsymbol{u}}_{\mathrm{l}} - c_l^2 \Delta \boldsymbol{u}_{\mathrm{l}}\right). \end{aligned} \tag{B.173}$$

We can follow a derivation similar to that of Eq. (B.172) and obtain

$$\ddot{\boldsymbol{u}}_{\mathrm{t}} - c_t^2 \Delta \boldsymbol{u}_{\mathrm{t}} = 0. \tag{B.174}$$

The equations of elastic waves are summarized as follows;

$$\ddot{\boldsymbol{u}}_{\mathrm{t}} = c_t \Delta \boldsymbol{u}_t \tag{B.175}$$
$$\ddot{\boldsymbol{u}}_{\mathrm{l}} = c_l \Delta \boldsymbol{u}_l \tag{B.176}$$
$$\mathrm{div}\,\boldsymbol{u}_t = 0 \tag{B.177}$$
$$\mathrm{rot}\,\boldsymbol{u}_l = 0 \tag{B.178}$$
$$\frac{c_t}{c_l} = \sqrt{\frac{1-2\nu}{2(1-\nu)}}. \tag{B.179}$$

Surface wave

From the Eqs. (B.175),(B.176), (B.177),(B.178),(B.179), we derive the surface elastic waves, or the Rayleigh waves, that propagate near the surface region. Let the medium be in $z < 0$ and the x axis is chosen as the propagation direction. We can suppose the following forms

$$u_{\mathrm{t}\alpha}(x,z) = u_{\mathrm{t}\alpha}^{(0)} e^{ik(x-c_s t)} e^{\kappa_{\mathrm{t}} z} \quad (\alpha = x,y,z) \tag{B.180}$$
$$u_{\mathrm{l}\alpha}(x,z) = u_{\mathrm{l}\alpha}^{(0)} e^{ik(x-c_s t)} e^{\kappa_{\mathrm{l}} z} \quad (\alpha = x,y,z). \tag{B.181}$$

with $c_s > 0$. Equation (B.175) gives

$$-c_s^2 k^2 = c_t^2(-k^2 + \kappa^2) \tag{B.182}$$

and thus

$$\kappa_{\mathrm{t}} = k\sqrt{1 - \left(\frac{c_s}{c_t}\right)^2}. \tag{B.183}$$

Similarly, Eq. (B.176) gives

$$\kappa_{\mathrm{l}} = k\sqrt{1 - \left(\frac{c_s}{c_l}\right)^2}. \tag{B.184}$$

For the decay property in the z direction, κ_{t} and κ_{l} should be positive ($\kappa_{\mathrm{t}}, \kappa_{\mathrm{l}} > 0$), which requires

$$c_s < c_t < c_l. \tag{B.185}$$

The boundary condition at the surface ($z = 0$) is written as

$$\hat{\sigma}\boldsymbol{n}_z = 0. \tag{B.186}$$

Here we define α as

$$\alpha \equiv \frac{c_s}{c_t}. \tag{B.187}$$

After several calculations [132], we obtain the following relations;

$$u_{\mathrm{t}y}^{(0)} = u_{\mathrm{l}y}^{(0)} = 0 \tag{B.188}$$

$$\begin{pmatrix} u_{\mathrm{t}x}^{(0)} \\ u_{\mathrm{t}z}^{(0)} \end{pmatrix} = a \begin{pmatrix} \kappa_{\mathrm{t}} \\ -ik \end{pmatrix}, \quad \begin{pmatrix} u_{\mathrm{l}x}^{(0)} \\ u_{\mathrm{l}z}^{(0)} \end{pmatrix} = b \begin{pmatrix} k \\ -i\kappa_{\mathrm{l}} \end{pmatrix} \tag{B.189}$$

$$\frac{a}{b} = -\frac{2-\alpha^2}{2\sqrt{1-\alpha^2}} \tag{B.190}$$

$$\alpha^6 - 8\alpha^4 + 8\left(3 - 2\frac{c_t^2}{c_l^2}\right)\alpha^2 - 16\left(1 - \frac{c_t^2}{c_l^2}\right) = 0. \tag{B.191}$$

Since the ratio c_t/c_l is a function of the Poisson ratio ν, as in Eq. (B.179), the parameter α is a function of the Poisson ratio ν ($\alpha = \alpha(\nu)$). Within the range of $0 \leq \nu \leq 1/2$, the numerical solution of Eq. (B.191) gives α uniquely as a monotonically increasing function of ν. The range of α is $0.874 \leq \alpha \leq 0.955$ within the above range of ν. It should be noted that the two components of surface wave, unlike those of bulk wave, are not independent, on account of the surface boundary condition of Eq. (B.186);

$$\frac{u_{\mathrm{t}x}^{(0)}}{u_{\mathrm{l}x}^{(0)}} = \frac{a\kappa_{\mathrm{t}}}{bk} = -\left(1 - \frac{c_s^2}{c_t^2}\right)^{1/2} \frac{2-\alpha^2}{2\sqrt{1-\alpha^2}} = -\frac{2-\alpha^2}{2} \tag{B.192}$$

$$\frac{u_{\mathrm{t}z}^{(0)}}{u_{\mathrm{l}z}^{(0)}} = \frac{ak}{b\kappa_{\mathrm{l}}} = -\left(1 - \frac{c_s^2}{c_l^2}\right)^{-1/2} \frac{2-\alpha^2}{2\sqrt{1-\alpha^2}}. \tag{B.193}$$

Appendix C

Continuum theory of fracture

C.1 Theory of elasticity in isotropic 2D medium

The theory of elasticity in isotropic 2D medium is briefly reviewed, particularly for describing the theory of fracture. See a text [137] as a reference.

As in the 3D medium (See Eq. (B.124)), the strain energy in an isotropic 2D medium is given as

$$U \equiv U_0 + \frac{\lambda_0}{2}\left(\mathrm{Tr}[u]\right)^2 + \mu_0 \sum_{ij} u_{ij}^2, \tag{C.1}$$

where the summation runs over the planer components (x, y). In this appendix, the components of the stress tensor are written as

$$\sigma_{ij} \Rightarrow \begin{pmatrix} \sigma_x & \tau_{xy} \\ \tau_{xy} & \sigma_y \end{pmatrix}. \tag{C.2}$$

The equation of balance is written in

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} = 0, \quad \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{xy}}{\partial x} = 0. \tag{C.3}$$

The condition of the compatibility (See Section B.3) is reduced to only one equation;

$$\frac{\partial^2 u_{xx}}{\partial y^2} + \frac{\partial^2 u_{xx}}{\partial y^2} - 2\frac{\partial^2 u_{xy}}{\partial x \partial y} = 0. \tag{C.4}$$

Stress and strain

The components of the stress are given, like Eq. (B.126), as

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \begin{pmatrix} \lambda_0 + 2\mu_0 & \lambda_0 & \\ \lambda_0 & \lambda_0 + 2\mu_0 & \\ & & 2\mu_0 \end{pmatrix} \begin{pmatrix} u_{xx} \\ u_{yy} \\ u_{xy} \end{pmatrix} \tag{C.5}$$

With a determinant

$$D \equiv \begin{vmatrix} \lambda_0 + 2\mu_0 & \lambda_0 \\ \lambda_0 & \lambda_0 + 2\mu_0 \end{vmatrix} = (\lambda_0 + 2\mu_0)^2 - \lambda_0^2 = 4\mu_0(\lambda_0 + \mu_0), \tag{C.6}$$

we obtain the components of the strain

$$u_{xx} = \frac{1}{D}\left\{(\lambda_0 + 2\mu_0)\sigma_x - \lambda_0 \sigma_y\right\} \tag{C.7}$$

$$u_{yy} = \frac{1}{D}\left\{(\lambda_0 + 2\mu_0)\sigma_y - \lambda_0 \sigma_x\right\} \tag{C.8}$$

$$u_{xy} = \frac{\tau_{xy}}{2\mu_0}. \tag{C.9}$$

When Eqs. (C.7),(C.8),(C.9) are substituted into Eq. (C.4), we obtain

$$\begin{aligned} 0 &= \frac{\partial^2 u_{xx}}{\partial y^2} + \frac{\partial^2 u_{xx}}{\partial y^2} - 2\frac{\partial^2 u_{xy}}{\partial x \partial y} \\ &= \frac{1}{D}\left[(\lambda_0 + 2\mu_0)\frac{\partial^2 \sigma_x}{\partial y^2} - \lambda_0 \frac{\partial^2 \sigma_y}{\partial y^2}\right. \\ &\quad \left. + (\lambda_0 + 2\mu_0)\frac{\partial^2 \sigma_y}{\partial x^2} - \lambda_0 \frac{\partial^2 \sigma_x}{\partial x^2} - \frac{D}{\mu_0}\frac{\partial^2 \tau_{xy}}{\partial x \partial y}\right]. \end{aligned} \tag{C.10}$$

Here Eq. (C.3) is used for eliminate τ_{xy} and obtain

$$\begin{aligned} 0 &= (\lambda_0+2\mu_0)\frac{\partial^2\sigma_x}{\partial y^2} - \lambda_0\frac{\partial^2\sigma_y}{\partial y^2} \\ &\quad +(\lambda_0+2\mu_0)\frac{\partial^2\sigma_y}{\partial x^2} - \lambda_0\frac{\partial^2\sigma_x}{\partial x^2} \\ &\quad -\frac{D}{2\mu_0}\left(\frac{\partial^2\sigma_x}{\partial x^2}+\frac{\partial^2\sigma_x}{\partial x^2}\right) \\ &= (\lambda_0+2\mu_0)\left\{\frac{\partial^2\sigma_x}{\partial y^2}+\frac{\partial^2\sigma_y}{\partial y^2}+\frac{\partial^2\sigma_y}{\partial x^2}+\frac{\partial^2\sigma_x}{\partial x^2}\right\} \\ &= (\lambda_0+2\mu_0)\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)(\sigma_x+\sigma_y), \end{aligned} \tag{C.11}$$

where we use the relation $D/(2\mu_0) = 2\lambda_0 + 2\mu_0$ that is derived from Eq. (C.6).

Stress function (1)

Equations (C.3) and (C.11) are the conditions on the stress components. The stress components can be expressed by a real function $A \equiv A(x,y)$ as

$$\sigma_x = \frac{\partial^2 A}{\partial y^2} \tag{C.12}$$

$$\sigma_y = \frac{\partial^2 A}{\partial x^2} \tag{C.13}$$

$$\tau_{xy} = -\frac{\partial^2 A}{\partial x \partial y}. \tag{C.14}$$

Equation (C.11) is reduced to the biharmonic property

$$\Delta^2 A = 0 \tag{C.15}$$

of the function A. This function is called 'Airy function'.

Here we transform the formula to those with the complex variable $z \equiv x+iy$. The two independent variables $(z,\bar{z})$ can be used instead of (x,y). Some mathematical relations are listed

$$\frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial z^2} + 2\frac{\partial^2}{\partial z \partial \bar{z}} + \frac{\partial^2}{\partial \bar{z}^2}, \tag{C.16}$$

$$\frac{\partial^2}{\partial y^2} = -\frac{\partial^2}{\partial z^2} + 2\frac{\partial^2}{\partial z \partial \bar{z}} - \frac{\partial^2}{\partial \bar{z}^2}, \tag{C.17}$$

$$\frac{\partial^2}{\partial x \partial y} = i\left(\frac{\partial^2}{\partial z^2} - \frac{\partial^2}{\partial \bar{z}^2}\right), \tag{C.18}$$

and so

$$4\frac{\partial^2}{\partial z \partial \bar{z}} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \tag{C.19}$$

$$4\frac{\partial^2}{\partial z^2} = \left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right) - 2i\frac{\partial^2}{\partial x \partial y}. \tag{C.20}$$

Using Eqs. (C.19,C.20), we can write

$$4\frac{\partial^2 A}{\partial z \partial \bar{z}} = \sigma_x + \sigma_y \tag{C.21}$$

$$4\frac{\partial^2 A}{\partial z^2} = (\sigma_x - \sigma_y) + 2i\tau_{xy} \tag{C.22}$$

Hereafter we denote differentiation or integration as

$$f'(z) \equiv \frac{df}{dz}, \quad f''(z) \equiv \frac{d^2 f}{dz^2},$$
$$f^{(-1)}(z) \equiv \int^z dz f(z), \quad f^{(-2)}(z) \equiv \int^z dz \int^z dz f(z).$$

In general, if a function f is holomorphic with respect to z, the relations

$$\frac{\partial f}{\partial \bar{z}} = 0 \tag{C.23}$$

$$\frac{\partial f}{\partial z} = f'(z) \tag{C.24}$$

are satisfied. The former one is the Cauchy-Riemann relation. The above relations are essential to the following discussions.

From Eq. (C.15) and Eq. (C.19), we obtain

$$\frac{\partial^4 A}{\partial z^2 \partial \bar{z}^2} = 0. \tag{C.25}$$

The function A is given by successive integrations of Eq. (C.25) with respect to z or $\bar{z}$;

$$\frac{\partial^3 A}{\partial z \partial \bar{z}^2} = f_1(\bar{z}) \tag{C.26}$$

$$\frac{\partial^2 A}{\partial z \partial \bar{z}} = f_1^{(-1)}(\bar{z}) + f_2(z) \tag{C.27}$$

$$\frac{\partial A}{\partial \bar{z}} = z f_1^{(-1)}(\bar{z}) + f_2^{(-1)}(z) + f_3(\bar{z}) \tag{C.28}$$

$$A = z f_1^{(-2)}(\bar{z}) + \bar{z} f_2^{(-1)}(z) + f_3^{(-1)}(\bar{z}) + f_4(z), \tag{C.29}$$

where f_1, f_2, f_3, f_4 are arbitrary holomorphic functions.

Since $A(x, y)$ must be real ($\bar{A} = A$), the above expression can be written as

$$\begin{aligned} A &= \frac{1}{2}\left(\bar{z}\phi(z) + z\overline{\phi(z)} + \chi(z) + \overline{\chi(z)}\right) \\ &= \mathrm{Re}\left[\bar{z}\phi(z) + \chi(z)\right], \end{aligned} \tag{C.30}$$

with holomorphic functions $\phi(z)$ and $\chi(z)$. The above functions ϕ and χ are called 'Goursat functions'.

Using Eqs. (C.21) and (C.22), one can obtain

$$\begin{aligned} \frac{\sigma_x + \sigma_y}{2} &= 2\frac{\partial^2 A}{\partial z \partial \bar{z}} \\ &= \left[\frac{\partial^2}{\partial z \partial \bar{z}}\left(\bar{z}\phi + \chi\right) + \mathrm{c.c}\right] \\ &= \left[\phi'(z) + \mathrm{c.c}\right] \\ &= 2\mathrm{Re}\left[\phi'\right] \end{aligned} \tag{C.31}$$

and

$$\begin{aligned}\frac{\sigma_x-\sigma_y}{2}+i\tau_{xy} &= \frac{\partial^2 A}{\partial z^2}\\ &= \frac{\partial^2}{\partial z^2}\left[\bar{z}\phi+\chi+z\bar{\phi}+\bar{\chi}\right]\\ &= \bar{z}\phi'+\chi''. \end{aligned} \tag{C.32}$$

Therefore, the components of stress are expressed by the Goursat functions as

$$\sigma_x = 2\mathrm{Re}\left[\phi'\right]-x\mathrm{Re}\left[\phi''\right]-y\mathrm{Im}\left[\phi''\right]-\mathrm{Re}\left[\chi''\right] \tag{C.33}$$

$$\sigma_y = 2\mathrm{Re}\left[\phi'\right]+x\mathrm{Re}\left[\phi''\right]+y\mathrm{Im}\left[\phi''\right]+\mathrm{Re}\left[\chi''\right] \tag{C.34}$$

$$\tau_{xy} = x\mathrm{Im}\left[\phi''\right]-y\mathrm{Re}\left[\phi''\right]+\mathrm{In}\left[\chi''\right]. \tag{C.35}$$

Stress function (2)

Other pairs of the stress functions can be used in stead of $\phi(z)$ and $\chi(z)$. One of such pairs is called 'Westergaard functions', denoted as $Z_I(z)$ and $Z_{II}(z)$. The definition is

$$Z_I = 2\phi'+z\phi''+\chi'' \tag{C.36}$$

$$iZ_{II} = z\phi''+\chi'' \tag{C.37}$$

or

$$2\phi = Z_I^{(-1)}-iZ_{II}^{(-1)} \tag{C.38}$$

$$2\chi' = \left\{Z_I^{(-1)}-zZ_I\right\}+i\left\{Z_{II}^{(-1)}+zZ_{II}\right\}. \tag{C.39}$$

Here Eq. (C.38) is directly derived from the integrations of Eqs. (C.36) and (C.37). Equation (C.38) is obtain, when we integrate Eq. (C.37);

$$\begin{aligned}2\chi' &= 2iZ_{II}^{(-1)}-2\int z\phi''dz\\ &= 2iZ_{II}^{(-1)}-2(z\phi'-\phi)\\ &= 2iZ_{II}^{(-1)}-2z\phi'+2\phi\\ &= 2iZ_{II}^{(-1)}-z(Z_I-iZ_{II})+(Z_I^{(-1)}-iZ_{II}^{(-1)})\\ &= \left\{Z_I^{(-1)}-zZ_I\right\}+i\left\{Z_{II}^{(-1)}+zZ_{II}\right\}, \end{aligned}$$

where the fourth equality is given by Eq. (C.38) and its derivative.

The components of the stress are given as

$$\begin{aligned}\sigma_x &= \mathrm{Re}[Z_I]-y\mathrm{Im}[Z_I']+2\mathrm{Im}[Z_{II}]-y\mathrm{Re}[Z_{II}']\\ \sigma_y &= \mathrm{Re}[Z_I]+y\mathrm{Im}[Z_I']+-y\mathrm{Re}[Z_{II}']\\ \tau_{xy} &= -y\mathrm{Re}[Z_I']+\mathrm{Re}[Z_{II}]-y\mathrm{Im}[Z_{II}'].\end{aligned} \tag{C.40}$$

Since the above forms will be reduced to simple ones on $y=0$, the Westergaard functions are useful for the crack analysis, in which the crack lies on $y=0$.

As a simple case, a uniform stress field

$$(\sigma_x, \sigma_y, \tau_{xy}) = (\sigma_x^{(0)}, \sigma_y^{(0)}, \tau_{xy}^{(0)}) \tag{C.41}$$

is expressed by the following constant functions

$$\begin{aligned} Z_I(z) &= \sigma_y^{(0)} \\ Z_{II}(z) &= \tau_{xy}^{(0)} + i\frac{\sigma_x^{(0)} - \sigma_y^{(0)}}{2}, \end{aligned} \tag{C.42}$$

where $\sigma_x^{(0)}$,$\sigma_y^{(0)}$ and $\tau_{xy}^{(0)}$ are real constants. Note that the imaginary constant term of Z_I does not contribute the stress fields.

Using the Westergaard functions, the Airy function, given as Eq. (C.30), is written as

$$\begin{aligned} A &= \mathrm{Re}\left[Z_I^{(-2)} - iyZ_I^{(-1)} - yZ_{II}^{(-1)}\right] \\ &= \mathrm{Re}\left[Z_I^{(-2)}\right] + y\mathrm{Im}\left[Z_I^{(-1)}\right] - y\mathrm{Re}\left[Z_{II}^{(-1)}\right]. \end{aligned} \tag{C.43}$$

The derivation of Eq. (C.43) is as follows; when we integrate Eq. (C.39), we obtain

$$\begin{aligned} \chi &= \frac{1}{2}\left[Z_I - \int zZ_I dz + iZ_{II} + i\int zZ_{II}dz\right] \\ &= \frac{1}{2}\left[Z_I^{(-2)} - (zZ_I^{(-1)} - Z_I^{(-2)}) + iZ_{II}^{(-2)} + i(zZ_{II}^{(-1)} - Z_{II}^{(-2)})\right] \\ &= Z_I^{(-2)} - \frac{1}{2}z(Z_I^{(-1)} - iZ_{II}^{(-1)}). \end{aligned} \tag{C.44}$$

On the other hand, Eq. (C.38) gives

$$\bar{z}\phi = \frac{1}{2}\bar{z}(Z_I^{(-1)} - iZ_{II}^{(-1)}). \tag{C.45}$$

When Eqs. (C.44) and (C.45) are substituted into Eq. (C.30), we obtain Eq. (C.43).

Stress field with a crack

Now we would like to show the stress function for the stress field with a crack. The crack lies on the finite line connecting $z = -c$ and $z = c$, where c is a positive value. Three polar axes are defined as

$$z = r_{\mathrm{O}}e^{i\theta_{\mathrm{O}}}, \quad z - c = r_{\mathrm{A}}e^{i\theta_{\mathrm{A}}}, \quad z + c = r_{\mathrm{B}}e^{i\theta_{\mathrm{B}}} \tag{C.46}$$

and the variables

$$\bar{\theta} \equiv \frac{\theta_{\mathrm{A}} + \theta_{\mathrm{B}}}{2}, \quad \bar{r} \equiv \sqrt{r_{\mathrm{A}}r_{\mathrm{B}}} \tag{C.47}$$

are also defined. All the angles should satisfy $-\pi < \theta_{\mathrm{O}}, \theta_{\mathrm{A}}, \theta_{\mathrm{B}}, \bar{\theta} < \pi$. The geometry is shown in Fig. C.1.

A stress function is proposed as

$$Z_I = \sigma_0\frac{z}{\sqrt{z^2 - c^2}}, \quad Z_{II} = 0. \tag{C.48}$$

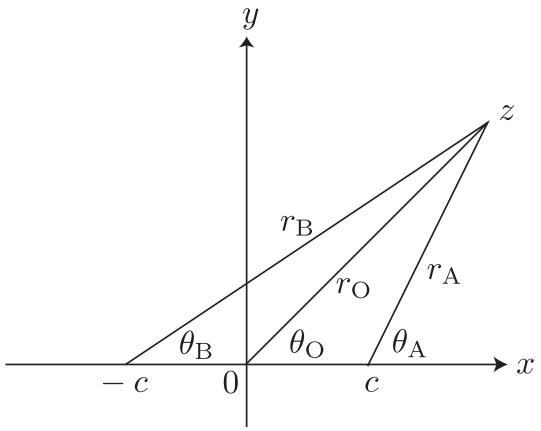


Figure C.1: Geometry of the three polar axes $(r_{\rm O}, \theta_{\rm O})$, $(r_{\rm A}, \theta_{\rm A})$, $(r_{\rm B}, \theta_{\rm B})$.

The branch cut is chosen at the crack region, that is, the line connecting $z = -c$ and $z = c$. This choice define the function $\sqrt{z^2 - c^2}$ as a single-value function without the crack region;

$$\sqrt{z^2 - c^2} = \sqrt{r_{\rm A}}\sqrt{r_{\rm B}} e^{i\theta_{\rm A}/2} e^{i\theta_{\rm B}/2} = \bar{r} e^{i\bar{\theta}}, \tag{C.49}$$

whose asymptotic behavior is

$$\sqrt{z^2 - c^2} \approx z, \quad (|z| \gg c). \tag{C.50}$$

Eq. (C.40) is reduced to, with the fact of $y = r_0 \sin\theta_O$,

$$\begin{aligned} \sigma_x &= \mathrm{Re}[Z_I] - r_O \sin\theta_O \,\mathrm{Im}[Z_I'] \\ \sigma_y &= \mathrm{Re}[Z_I] + r_O \sin\theta_O \,\mathrm{Im}[Z_I'] \\ \tau_{xy} &= -r_O \sin\theta_O \,\mathrm{Re}[Z_I']. \end{aligned} \tag{C.51}$$

Hereafter we show that the function in Eq. (C.48) gives the stress field of a crack under an external uniform stress field. As an asymptotic behavior, the stress function is reduced to

$$Z_I = \sigma_0, \quad Z_I' = 0 \quad \text{at} \quad |z| \gg c, \tag{C.52}$$

which gives the uniform stress field of

$$(\sigma_x, \sigma_y, \tau_{xy}) = (\sigma_0, \sigma_0, 0) \quad \text{at} \quad |z| \gg c. \tag{C.53}$$

Using the above polar axes, we obtain

$$Z_I = \sigma_0 \frac{z}{(z^2 - c^2)^{1/2}} = \sigma_0 \frac{r_{\rm O}}{\bar{r}} e^{i(\theta_{\rm O} - \bar{\theta})} \tag{C.54}$$

and

$$\begin{aligned} Z_I' &= \sigma_0 \frac{1}{(z^2 - c^2)^{1/2}} - \frac{1}{2} \frac{2z^2}{(z^2 - c^2)^{3/2}} \\ &= \sigma_0 \frac{1}{(z^2 - c^2)^{1/2}} \left(1 - \frac{z^2}{z^2 - c^2}\right) \\ &= \sigma_0 \frac{-c^2}{(z^2 - c^2)^{3/2}} = \sigma_0 \frac{-c^2}{\bar{r}^3} e^{-3i\bar{\theta}}. \end{aligned} \tag{C.55}$$

or

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \sigma_0 \frac{r_{\rm O}}{\bar{r}} \begin{pmatrix} \cos(\theta_{\rm O} - \bar{\theta}) - \frac{c^2}{\bar{r}^2} \sin\theta_{\rm O} \sin 3\bar{\theta} \\ \cos(\theta_{\rm O} - \bar{\theta}) + \frac{c^2}{\bar{r}^2} \sin\theta_{\rm O} \sin 3\bar{\theta} \\ \frac{c^2}{\bar{r}^2} \sin\theta_{\rm O} \cos 3\bar{\theta} \end{pmatrix} \tag{C.56}$$

Eq. (C.56) satisfy the free surface boundary

$$\sigma_y = 0, \quad \tau_{xy} = 0 \quad \text{at} \quad y = 0, -c < x < c \tag{C.57}$$

because of

$$\theta_{\rm O} = 0, \quad \bar{\theta} = \pi/2 \quad \text{at} \quad y = 0. \tag{C.58}$$

From the above analysis, we can conclude that the stress function in Eq. (C.56) shows a situation of a crack under the external uniform stress field in Eq. (C.53).

For the situation of the opening mode or 'mode I', the stress field should satisfy

$$(\sigma_x, \sigma_y, \tau_{xy}) = (0, \sigma_0, 0) \quad \text{at} \quad |z| \gg c, \tag{C.59}$$

not Eq. (C.53). The desirable stress functions are given by

$$\begin{aligned} Z_I &= \sigma_0 \frac{z}{\sqrt{z^2 - c^2}} \\ Z_{II} &= -i\frac{\sigma_0}{2}. \end{aligned} \tag{C.60}$$

The above functions can be decomposed into two terms. One gives the stress field of Eq. (C.56) and the other gives the uniform stress field of Eq. (C.41) with the choice of $(\sigma_x^{(0)}, \sigma_y^{(0)}, \tau_{xy}^{(0)}) = (-\sigma_0, 0, 0)$. The resultant stress field is the sum of that of Eq. (C.56) and

$$(\sigma_x, \sigma_y, \tau_{xy}) = (-\sigma_0, 0, 0), \tag{C.61}$$

which satisfies the boundary conditions Eqs. (C.57) and (C.59).

Now we show the stress field near the crack tip ($z \approx c$). We redefine the notation

$$r_{\rm A} \Rightarrow r, \quad \theta_{\rm A} \Rightarrow \theta \quad \sigma_0 \Rightarrow \frac{K}{\sqrt{\pi c}}. \tag{C.62}$$

In the region near the crack tip ($r \ll c$), the expressions

$$r_{\rm B} \approx 2c, \quad \theta_{\rm B} \approx 0 \tag{C.63}$$

are obtained, which results in

$$\bar{r} \approx \sqrt{2cr}, \quad \bar{\theta} \approx \frac{\theta}{2}. \tag{C.64}$$

The expressions for $(r_{\rm O}, \theta_{\rm O})$ should be based on the relations

$$y = r_{\rm O} \sin\theta_{\rm O} = r \sin\theta \tag{C.65}$$

$$x = r_{\rm O} \cos\theta_{\rm O} = r \cos\theta \tag{C.66}$$

$$r_{\rm O} \approx c, \tag{C.67}$$

which results in

$$\sin\theta_{\rm O} = \frac{r}{r_{\rm O}}\sin\theta \approx \frac{r}{c}\sin\theta \tag{C.68}$$

$$\cos\theta_{\rm O} = \frac{1}{r_{\rm O}}\left(c + r\cos\theta\right) \approx \frac{c}{r_{\rm O}} \approx 1. \tag{C.69}$$

Using the above expressions, the essential terms in Eq. (C.56) are calculated

$$\begin{aligned}
\cos(\theta_{\rm O} - \bar{\theta}) &= \cos\theta_{\rm O} \times \cos\bar{\theta} - \sin\theta_{\rm O} \times \sin\bar{\theta} \\
&\approx 1 \times \cos\frac{\theta}{2} - \frac{r}{c}\sin\theta \times \sin\frac{\theta}{2} \approx \cos\frac{\theta}{2} \qquad \text{(C.70)} \\
\frac{c^2}{\bar{r}^2} \times \sin\theta_{\rm O} &\approx \frac{c^2}{2cr} \times \frac{r}{c}\sin\theta \\
&= \frac{1}{2}\sin\theta = \cos\frac{\theta}{2}\sin\frac{\theta}{2} \qquad \text{(C.71)} \\
\sigma_0 \times \frac{r_{\rm O}}{\bar{r}} &= \frac{K}{\sqrt{\pi c}} \times \frac{c}{\sqrt{2cr}} = \frac{K}{\sqrt{2\pi r}}. \qquad \text{(C.72)}
\end{aligned}$$

When the above expressions are substituted into Eq. (C.56), we obtain

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} \approx \frac{K}{\sqrt{2\pi r}}\cos\frac{\theta}{2}\begin{pmatrix} 1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2} \\ 1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2} \\ \sin\frac{\theta}{2}\cos\frac{3\theta}{2} \end{pmatrix}, \tag{C.73}$$

which diverges at the crack tip ($r = 0$). Since the stress field in Eq. (C.61), a constant field, does not diverge, the asymptotic stress field near the crack tip is given by Eq. (C.73) for the 'mode I' crack. The factor K is usually called 'stress intensity factor' for the 'mode I' crack.

C.2 Theories of fracture

In this section, the continuum theory for brittle fracture is briefly reviewed. For the overview, see references [90, 91, 92, 93]. Usually three basic 'modes' of crack-surface displacement are distinguished, which are schematically shown in Fig. C.2. Among them, we pick out the 'mode I' or the opening mode, which is the most important mode for the fracture propagation in highly brittle solids. Hereafter we consider the situation shown in Fig.C.3(a), in which the uniaxial external load σ is imposed on a sample with a crack. The crack length is denoted as $2c$.

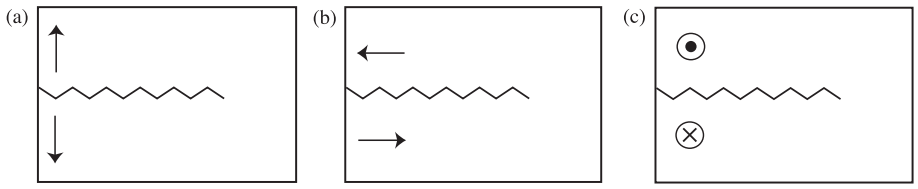


Figure C.2: Schematic pictures of the three fracture modes. (a) mode I or the opening mode, (b) mode II or the sliding mode, (c) mode III or the tearing mode. The arrow indicates the direction of displacements. In (c), the displacements of the upper and lower peaces are perpendicular to the paper in the opposite directions.

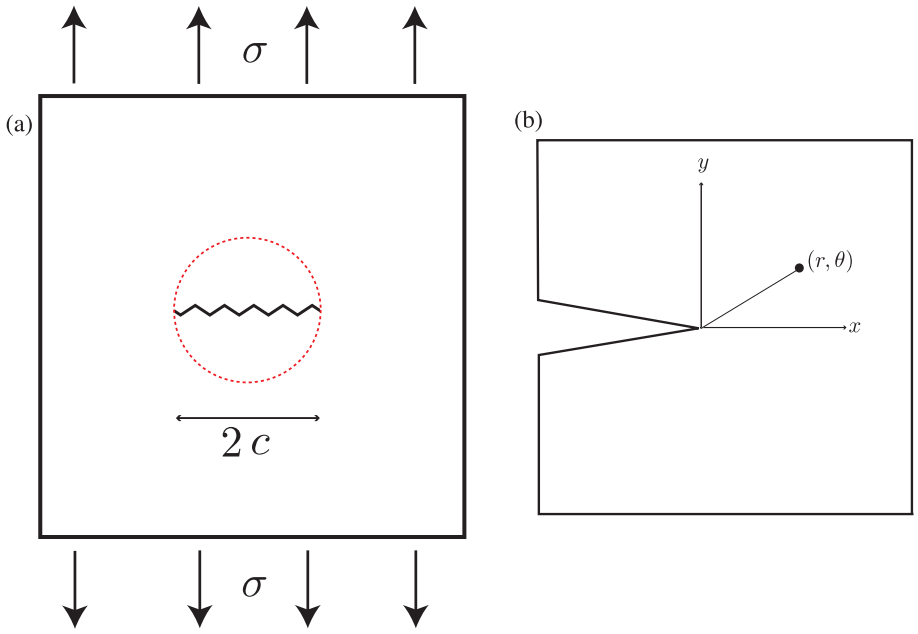


Figure C.3: (a) Schematic picture of a sample with a crack under the external load. The length of the crack is defined as $2c$. The red circle is just a eye guide for a circular area with the radius of c. (b) The magnification of the region near the crack tip ($r \ll c$).

Asymptotic stress field

The first fundamental concept for fracture is the asymptotic stress field at the crack tip [138]. The two dimensional polar axis (r, θ) is used $(x = r\cos\theta, y = r\sin\theta)$, of which origin is chosen at the crack tip, as in Fig.C.3(b). With the above axis, the stress tensor elements $\sigma_{xx}, \sigma_{xy}(= \sigma_{yx}), \sigma_{yy}$ are expressed near the crack tip $(r \ll c)$ as

$$\sigma_{ij}(r, \theta) = \frac{K}{\sqrt{2\pi r}} f_{ij}(\theta). \tag{C.74}$$

$$K \equiv \sigma\sqrt{\pi c}, \tag{C.75}$$

where $f_{ij}(\theta)$ is proper functions and σ is the external stress. K is called 'stress intensity factor'. The derivation of Eq. (C.74) is given in Appendix C.1 with the explicit forms of $f_{ij}(\theta)$. Here the fracture toughness, denoted as $K_{\rm c}$, is defined as the critical value of K for the fracture propagation. The fracture toughness $K_{\rm c}$ can be observed experimentally.

The stress field given in Eq. (C.74) diverges at the crack tip $(r = 0)$. This divergence originates in the large deformation at the region near the crack tip, which is beyond the theory of linear elasticity. Several energetic descriptions beyond linear elasticity were proposed for describing the local region near the crack tip. The size and the atomistic picture of such a local region may be different among materials. See the textbooks listed in the beginning of this section.

Griffith theory

The second work is the Griffith theory [89], in which an energy competition is essential within the total energy of

$$U = -\frac{\pi c^2 \sigma^2}{E} L_z + 4\gamma c L_z. \tag{C.76}$$

The first term is the energy gain of the strain relaxation, while the second term is the loss of the surface formation energy. Here L_z and E are the sample thickness and the Young modulus. The present Young modulus E is one within a two-dimensional case and corresponds to $E = E_{\rm 3D}$ in plane stress ('thin' plates) or $E = E_{\rm 3D}/(1-\nu^2)$ in plane strain ('thick' plates), with the ordinary Young modulus $E_{\rm 3D}$ and the Poisson ratio ν. The quantity γ is the loss of the surface formation energy per unit surface area. The factor πc^2 in the first term is the area of the dashed circle in Fig.C.3(a), of which radius is given by c. The factor appears, when one assumes that the crack with the length of $2c$ releases the strain energy within the circular area of the dashed circle in in Fig.C.3(a). The factor $4c$ in the second term originates from the fact that the two cleavage surfaces are formed as the upper and lower surfaces of the crack plane and each surface has the length of $2c$.

The total energy U shows a peak at

$$0 = \frac{dU}{dc} = -\frac{\pi\sigma^2}{E} 2c L_z + 4\gamma L_z, \tag{C.77}$$

which gives a length quantity $c_{\rm G}$ as

$$c = c_{\rm G} \equiv \frac{2}{\pi}\frac{\gamma E}{\sigma^2}. \tag{C.78}$$

If the external load σ is given, the value of the length $c_{\rm G}$ is determined uniquely. The length $c_{\rm G}$ gives the critical crack length for the spontaneous fracture propagation. The essence is the fact that the first term in Eq. (C.76) is a volume term that is proportional to (length)^3, while the second term is a surface term that is proportional to (length)^2. Therefore, dimensional analysis gives a typical length scale $c_{\rm G}$. This energy competition between volume and surface terms is analogous to the theory of nucleation, which is seen in textbooks of statistical mechanics [101].

With the definition of the fracture toughness $K_{\rm c}$, Eq. (C.75) and Eq. (C.78) give an important relation

$$(c_{\rm G} =)\frac{K_{\rm c}^2}{\sigma^2\pi} = \frac{2}{\pi}\frac{\gamma E}{\sigma^2}, \tag{C.79}$$

or

$$K_{\rm c}^2 = 2\gamma E. \tag{C.80}$$

Equation (C.80) shows the direct relation between the fracture toughness $K_{\rm c}$ and the surface formation energy γ.

The validity of Eq. (C.80) was investigated in the cleaved Si(111) surface [102]. In results, the value of γ is estimated to be $\gamma \approx 1.1[{\rm J/m^2}]$ within the factor of two [102]. Equation (C.80) with the above value of γ gives a consistent explanation between the results of electronic structure calculations and several experimental results. The corresponding atomistic picture was discussed in Section 6.1.

Since the essence of the Griffith theory is dimensional analysis, the energy term 4 γc is essential only in its linear dependence on the crack length. In other words, the factor γ can be generally defined as a 'dissipative' energy that is required in the fracture propagation by the unit length. For such generalizations and the corresponding atomistic picture, see the textbooks listed in the beginning of this section.

Mott theory

Third concept is the Mott theory [139], in which the kinetic energy term K is introduced in the total energy. In fracture dynamics, the release of the strain energy should be transformed into the local kinetic energy of atoms, as well as the surface formation energy. The explicit form of K is given by

$$K = \frac{\rho}{2}kv^2c^2\frac{\sigma^2}{E^2}L_z, \tag{C.81}$$

where v is the crack propagation speed. ρ is the density and k is a numerical factor. The dependence of v on c is now ignored for a stationary solution (v = (const)). Using the kinetic energy of Eq. (C.81), the speed v is determined by

$$\frac{d}{dc}(U+K) = 0. \tag{C.82}$$

Now the static energy per unit thickness (U/L_z) is rewritten as

$$\begin{aligned}\frac{U}{L_z} &= -\frac{\pi c^2\sigma^2}{E} + 4\gamma c \\ &= -\frac{\pi\sigma^2}{E}(c-c_{\rm G})^2 + \frac{\pi\sigma^2}{E}c_{\rm G}^2.\end{aligned} \tag{C.83}$$

with the critical crack length $c_{\rm G}$. Equation (C.82) is calculated as

$$\begin{aligned} 0 &= \frac{1}{L_z}\frac{d}{dc}(U+K) \\ &= \frac{d}{dc}\left(-\frac{\pi\sigma^2}{E}(c-c_{\rm G})^2+\frac{\pi\sigma^2}{E}c_{\rm G}^2+\frac{\rho}{2}kv^2c^2\frac{\sigma^2}{E^2}\right) \\ &= -\frac{2\pi\sigma^2}{E}(c-c_{\rm G})+k\rho v^2 c\frac{\sigma^2}{E^2}, \end{aligned} \tag{C.84}$$

or

$$v=\sqrt{\frac{2\pi}{k}}\sqrt{\frac{E}{\rho}}\left(1-\frac{c_{\rm G}}{c}\right)^{1/2}. \tag{C.85}$$

With increasing the crack length ($c \gg c_{\rm G}$), the value of v in Eq. (C.85) will reach the stationary solution of

$$v \to \sqrt{\frac{2\pi}{k}}\sqrt{\frac{E}{\rho}} \quad (c \gg c_{\rm G}). \tag{C.86}$$

Here we note that $\sqrt{E/\rho}$ is the elastic wave speed. Later theoretical works [140, 141, 91] predict that the crack propagation speed can not exceed the Rayleigh wave speed, which is seen experimentally. Since the above prediction is within continuum mechanics, its validity is sometimes focused in atomistic pictures, such as Ref.[142].

Appendix D

Miscellaneous notes

D.1 Conventional Wannier state in one-dimensional system

In this appendix, we focus on the conventional Wannier state, not the *generalized* Wannier state explained in Section 2.3.

Suppose a one-dimensional system with an isolated band, in which the Schrödinger equation has the solution of Bloch states;

$$H|\psi_k\rangle = \varepsilon(k)|\psi_k\rangle. \tag{D.1}$$

The length of the unit cell is denoted as a. Since the Bloch states is periodic in reciprocal space

$$\psi_{k+2\pi/a}(x) = \psi_k(x), \tag{D.2}$$

they can be expanded within Fourier series;

$$\psi_k(x) = \sum_{l=-\infty}^{\infty} W_l(x) e^{ilka}. \tag{D.3}$$

Similarly, the dispersion curve $\varepsilon(k)$ can be also expanded with the Fourier series;

$$\varepsilon(k) = \sum_{l=-\infty}^{\infty} \varepsilon^{(\mathrm{W})}(l) e^{ilka}. \tag{D.4}$$

Due of the mathematical relation

$$\int_{-\pi}^{\pi} e^{i(l-l')\theta} d\theta = 2\pi\delta_{l,l'}, \tag{D.5}$$

Eq. (D.3) gives

$$W_l(x) = \int_{-\pi/a}^{\pi/a} \frac{dk}{(2\pi/a)} e^{-ilka} \psi_k(x), \tag{D.6}$$

which correspond to the definition of the conventional (isolated-band) Wannier state. Note that the region $-\pi/a < k < \pi/a$ is the first Brillouin zone and the factor $2\pi/a$ is its volume ($\int_{(1.\mathrm{B.Z.})} dk = 2\pi/a$).

If we define the length of $L \equiv aN$ with an integer N, the inner product between the Wannier states can be defined as

$$\begin{aligned} \langle W_l | W_{l'} \rangle_L &\equiv \int_0^L W_l(x) W_{l'}(x) dx \\ &= \int_{-\pi/a}^{\pi/a} \frac{dk}{(2\pi/a)} \int_{-\pi/a}^{\pi/a} \frac{dk'}{(2\pi/a)} e^{ilka} e^{-il'k'a} \langle \psi_k | \psi_{k'} \rangle_L. \end{aligned} \tag{D.7}$$

If we impose the normalization condition on the Wannier state

$$\langle W_l | W_l \rangle_L = 1, \tag{D.8}$$

the Bloch states should satisfy the following condition

$$\langle \psi_k | \psi_{k'} \rangle_L = \frac{2\pi}{a} \delta(k - k'). \tag{D.9}$$

With respect to the Wannier states, the Hamiltonian matrix is given as

$$\begin{aligned}
\langle W_l|H|W_{l'}\rangle_L &= \int_{-\pi/a}^{\pi/a}\frac{dk}{(2\pi/a)}\int_{-\pi/a}^{\pi/a}\frac{dk'}{(2\pi/a)}e^{ilka}e^{-il'ka}\langle\psi_k|H|\psi_k\rangle_L \\
&= \int_{-\pi/a}^{\pi/a}\frac{dk}{(2\pi/a)}\int_{-\pi/a}^{\pi/a}\frac{dk'}{(2\pi/a)}e^{ilka}e^{-il'ka}\frac{2\pi}{a}\varepsilon(k)\delta(k-k') \\
&= \int_{-\pi/a}^{\pi/a}\frac{dk}{(2\pi/a)}e^{i(l-l')ka}\varepsilon(k) \\
&= \varepsilon^{(\mathrm{W})}(l-l'), \qquad \text{(D.10)}
\end{aligned}$$

where the last equality is given by Eq. (D.4). In short, the off-diagonal Hamiltonian matrix with respect to the Wannier state is given by the Fourier coefficients of the dispersion curve $\varepsilon(k)$.

D.2 Density matrix in free electron system

In this appendix, we derive several equations given in Section 2.4. All the notations are the same as in Section 2.4;

(i) Equation (2.54) is derived as follows;

$$\begin{aligned}\frac{E}{V} &= \int_{k<k_{\rm F}} \frac{d\boldsymbol{k}}{(2\pi)^3}\frac{1}{2}k^2 \\ &= 4\pi\int_0^{k_{\rm F}} k^2\frac{1}{2}k^2\frac{dk}{(2\pi)^3} \\ &= \frac{1}{(2\pi)^2}\int_0^{k_{\rm F}} k^4 dk \\ &= \frac{1}{(2\pi)^2}\frac{k_{\rm F}^5}{5} = \frac{k_{\rm F}^5}{20\pi^2}. \end{aligned} \tag{D.11}$$

(ii) Equation (2.56) is derived as follows;

$$\begin{aligned}\rho(r) &\equiv V\int\frac{d\boldsymbol{k}}{(2\pi)^3}\frac{1}{V}e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)} \\ &= \frac{1}{(2\pi)^3}\int_0^{k_{\rm F}} k^2 dk(2\pi)\int_{-1}^{1} dt e^{ikrt} \\ &= \frac{1}{(2\pi)^2}\int_0^{k_{\rm F}} k^2\left[\frac{e^{ikrt}}{ikrt}\right]_{t=-1}^{t=1} dk \\ &= \frac{1}{(2\pi)^2}\int_0^{k_{\rm F}} k^2\frac{e^{ikr}-e^{-ikr}}{ikr}dk \\ &= \frac{1}{(2\pi)^2}\int_0^{k_{\rm F}} k^2 2\frac{\sin kr}{kr}dk \\ &= \frac{2}{(2\pi)^2 r}\int_0^{k_{\rm F}} k\sin kr dk \\ &= \frac{2}{(2\pi)^2 r}\left\{\left[-\frac{k\cos kr}{r}\right]_0^{k_{\rm F}} + \frac{1}{r}\int_0^{k_{\rm F}}\cos kr dk\right\} \\ &= \frac{2}{(2\pi)^2 r}\left\{\left[-\frac{k\cos kr}{r}\right]_0^{k_{\rm F}} + \left[\frac{\sin kr}{r^2}\right]_0^{k_{\rm F}}\right\} \\ &= \frac{2}{(2\pi)^2}\left\{-\frac{k_{\rm F}}{r^2}\cos k_{\rm F}r + \frac{1}{r^3}\sin k_{\rm F}r\right\}. \end{aligned} \tag{D.12}$$

(iii) Equation (2.58) is derived as follows;

$$\begin{aligned}\rho(r) &\propto -\frac{k_{\rm F}}{r^2}\cos k_{\rm F}r + \frac{1}{r^3}\sin k_{\rm F}r \\ &= -\frac{k_{\rm F}}{r^2}\left[1-\frac{k_{\rm F}^2}{2}r^2 + \frac{k_{\rm F}^4}{24}r^4 + O(r^6)\right] \\ &\quad +\frac{1}{r^3}\left[k_{\rm F}r - \frac{k_{\rm F}^3}{6}r^3 + \frac{k_{\rm F}^5}{120}r^5 + O(r^7)\right] \end{aligned}$$

$$
\begin{aligned}
&= (-k_{\rm F}+k_{\rm F})\,\frac{1}{r^2}+\left(\frac{k_{\rm F}^3}{2}-\frac{k_{\rm F}^3}{6}\right)+\left(\frac{-k_{\rm F}^5}{24}+\frac{k_{\rm F}^5}{120}\right)r^2+O(r^4)\\
&= \frac{k_{\rm F}^3}{6}-\frac{k_{\rm F}^5}{30}r^2+O(r^4)\\
&= C_0-\frac{C_2}{2}r^2+O(r^4), \qquad \text{(D.13)}
\end{aligned}
$$

where the last equality is obtained with the notations of Eqs. (2.59).

(iv) Using the relation

$$
r^2\frac{d\rho}{dr}=\frac{2}{(2\pi)^2}\left(-C_2r^3+O(r^5)\right), \qquad \text{(D.14)}
$$

Eq. (2.61) is derived as follows;

$$
\begin{aligned}
E_{\rm sphere}(\varepsilon) &\equiv \int_{r<\varepsilon} d\boldsymbol{r}\frac{-\Delta_{\boldsymbol{r}}}{2}\rho_{\rm GS}(r)\\
&= \frac{-1}{2}4\pi\int_0^{\varepsilon} r^2dr\frac{1}{r^2}\left(\frac{d}{dr}r^2\frac{d\rho}{dr}\right)\\
&= -(2\pi)\left[r^2\frac{d\rho}{dr}\right]_{r=0}^{r=\varepsilon}\\
&= -(2\pi)\frac{2}{(2\pi)^2}\left[-C_2\varepsilon^3\right]+O(\varepsilon^5)\\
&= \frac{C_2}{\pi}\varepsilon^3+O(\varepsilon^5). \qquad \text{(D.15)}
\end{aligned}
$$

D.3 Lanczos method

This appendix is a brief review of the Lanczos method [143], which gives the foundation of the recursion method [51] (See Section 2.5). The Lanczos method is also used in the variational order-N method (See Section 5.2).

Suppose an Hamiltonian operator $\hat{H}$ or an explicit $M \times M$ Hamiltonian matrix H. From an 'input' (normalized) vector $|u\rangle$, we can construct a set of vectors

$$|u\rangle, \quad H|u\rangle, \quad H^2|u\rangle, \quad ..., \quad H^{(N-1)}|u\rangle, \tag{D.16}$$

where N is an integer ($N \leq M$). In general, the above set of vectors contains N independent freedoms. Though the above vectors are not orthogonal, we can transform them into a orthogonal vectors, as follows; If a vector $|v\rangle$ is defined as

$$|v\rangle \equiv (1 - |u\rangle\langle u|)H|u\rangle, \tag{D.17}$$

the vector $|v\rangle$ is orthogonal to the vector $|u\rangle$;

$$\begin{aligned} \langle u|v\rangle &= \langle u|\{1 - |u\rangle\langle u|\}H|u\rangle \\ &= \langle u|H|u\rangle - \langle u|u\rangle\langle u|H|u\rangle = 0 \end{aligned} \tag{D.18}$$

Based on the above fact, an orthogonal basis set $\{|u_j\rangle\}_j$ ($j = 1, 2....., N$) can be constructed;

$$|u_1\rangle \equiv |u\rangle \tag{D.19}$$

$$b_1|u_2\rangle = H|u_1\rangle - a_1|u_1\rangle \tag{D.20}$$

$$b_n|u_{n+1}\rangle = H|u_n\rangle - a_n|u_n\rangle - b^*_{n-1}|u_{n-1}\rangle \tag{D.21}$$

where

$$a_n \equiv \langle u_n|H|u_n\rangle \tag{D.22}$$

$$b_n \equiv \langle u_{n+1}|H|u_n\rangle \tag{D.23}$$

Using the above recurrence relation, the vectors are successively generated

$$|u_1\rangle \Rightarrow |u_2\rangle \Rightarrow |u_3\rangle \Rightarrow |u_4\rangle...., \tag{D.24}$$

and the set of resultant vectors satisfy the orthogonal relation

$$\langle u_i|u\rangle_j = \delta_{ij}. \tag{D.25}$$

In the case of $N = M$, the resultant set of the vector $\{\boldsymbol{u}_i\}$ forms a $M \times M$ unitary matrix U as

$$U \equiv (\boldsymbol{u}_1\, \boldsymbol{u}_2\, \boldsymbol{u}_3 ... \boldsymbol{u}_M\,). \tag{D.26}$$

This procedure is, formally, one of the tridiagonalization procedure of H;

$$U^{-1}HU = \begin{pmatrix} a_1 & b_1 & & & & \\ b_1^* & a_2 & b_2 & & & \\ & b_2^* & a_3 & b_3 & & \\ & & ... & ... & ... & \\ & & & b^*_{M-1} & a_{M-1} & b_M \\ & & & & b^*_M & a_M \end{pmatrix}. \tag{D.27}$$

Hereafter the discussion is restricted, for simplicity, to the case with real variables ($b_i^* = b_i$). With $N < M$, an operator $\hat{H}'$ is defined as

$$\hat{H}' \equiv \sum_{i,j}^{N} |u_i\rangle H'_{ij} \langle u_i| \tag{D.28}$$

with the matrix H'_{ij} of

$$H'_{ij} \equiv \begin{pmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & b_2 & a_3 & b_3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & b_{N-2} & a_{N-1} & b_{N-1} \\ & & & & b_{N-1} & a_N \end{pmatrix}_{ij} \tag{D.29}$$

The operator $\hat{H}'$ satisfies

$$\hat{H}'|u_i\rangle = \hat{H}|u_i\rangle \quad (i = 1, 2, ...N-1). \tag{D.30}$$

In the Lanczos method, the matrix H' is discussed, instead of H. The eigen value problem of $\hat{H}'$ is given as

$$\begin{pmatrix} \varepsilon - a_1 & -b_1 & & & & \\ -b_1 & \varepsilon - a_2 & -b_2 & & & \\ & -b_2 & \varepsilon - a_3 & -b_3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & -b_{N-2} & \varepsilon - a_{N-1} & -b_{N-1} \\ & & & & -b_{N-1} & \varepsilon - a_N \end{pmatrix} \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ \cdots \\ P_{N-2} \\ P_{N-1} \end{pmatrix} = 0. \tag{D.31}$$

Let the function $\Delta_n(\varepsilon)$ represent the determinant of the partial matrix that contains only the first n rows and columns of the tridiagonal matrix in Eq. (D.31). The eigen values $\{\varepsilon_\alpha\}$ are given by the zeros of $\Delta_N(\varepsilon)$;

$$\Delta_N(\varepsilon) = 0 \leftrightarrow \varepsilon = \varepsilon_0, \varepsilon_1, ..., \varepsilon_{N-1} \tag{D.32}$$

The functions $\{\Delta_1(\varepsilon), \Delta_2(\varepsilon)...\}$ has the recurrence relation of

$$\Delta_{n+1}(\varepsilon) = (\varepsilon - a_n)\Delta_n(\varepsilon) - b_n^2 \Delta_{n-1}(\varepsilon), \tag{D.33}$$

which is proved by the Laplace expansion. For example, Eq. (D.33) with $n = 3$ is given by the following calculations

$$\begin{aligned} & \begin{vmatrix} \varepsilon - a_0 & -b_1 & & \\ -b_1 & \varepsilon - a_1 & -b_2 & \\ & -b_2 & \varepsilon - a_2 & -b_3 \\ & & -b_3 & \varepsilon - a_3 \end{vmatrix} \\ = & (\varepsilon - a_3) \begin{vmatrix} \varepsilon - a_0 & -b_1 & \\ -b_1 & \varepsilon - a_1 & -b_2 \\ & -b_2 & \varepsilon - a_2 \end{vmatrix} + b_3 \begin{vmatrix} \varepsilon - a_0 & -b_1 & \\ -b_1 & \varepsilon - a_1 & -b_2 \\ & & -b_3 \end{vmatrix} \\ = & (\varepsilon - a_3) \begin{vmatrix} \varepsilon - a_0 & -b_1 & \\ -b_1 & \varepsilon - a_1 & -b_2 \\ & -b_2 & \varepsilon - a_2 \end{vmatrix} - b_3^2 \begin{vmatrix} \varepsilon - a_0 & -b_1 \\ -b_1 & \varepsilon - a_1 \end{vmatrix}. \end{aligned} \tag{D.34}$$

If we define formally $\Delta_0(\varepsilon), \Delta_{-1}(\varepsilon)$ as

$$\Delta_0(\varepsilon) \equiv 1, \quad \Delta_{-1}(\varepsilon) \equiv 0, \tag{D.35}$$

Eq. (D.33) is satisfied for $n \geq 0$,

For each eigen value ε_α, the eigen vector $(P_0, P_2, P_3....)$ is given by

$$(\varepsilon_\alpha - a_n)P_n(\varepsilon_\alpha) = b_{n+1}P_{n+1}(\varepsilon_\alpha) + b_{n-1}P_{n-1}(\varepsilon_\alpha), \tag{D.36}$$

because of Eq. (D.33). On the other hand, we can define, generally, $\{P_n(\varepsilon)\}$ as

$$(\varepsilon - a_n)P_n(\varepsilon) = b_{n+1}P_{n+1}(\varepsilon) + b_{n-1}P_{n-1}(\varepsilon) \tag{D.37}$$

for any value of ε. With the additional definition of

$$P_{-1}(\varepsilon) \equiv 0, \quad P_0(\varepsilon) \equiv 1, \tag{D.38}$$

the function $P_n(\varepsilon)$ is determined uniquely as the n-th order polynomials of ε. When Eqs. (D.33), (D.35) are compared with Eqs. (D.37), (D.38), we obtain

$$\Delta_n(\varepsilon_\alpha) = b_1 b_2 b_3 ... b_n P_n(\varepsilon). \tag{D.39}$$

In other words, the polynomial $P_n(\varepsilon)$ is proportional to the determinant $\Delta_n(\varepsilon_\alpha)$. An eigen vector

$$\hat{H}'|w_\alpha\rangle = \varepsilon_\alpha |w_\alpha\rangle \tag{D.40}$$

is given by

$$\begin{aligned} |w_\alpha\rangle &= \sum_{n=0}^{N-1} P_n(\varepsilon_\alpha)|u_n\rangle \\ &= |u\rangle + P_1(\varepsilon_\alpha)|u_1\rangle + P_2(\varepsilon_\alpha)|u_2\rangle + ... \end{aligned} \tag{D.41}$$

This eigen vector is not normalized ($\langle w_\alpha | w_\alpha \rangle \neq 1$) but has a property

$$\langle u | w_\alpha \rangle = 1. \tag{D.42}$$

With the norm

$$|w_\alpha| \equiv \sqrt{\langle w_\alpha | w_\alpha \rangle}, \tag{D.43}$$

the normalized eigen vectors $\{\, |w_\alpha|^{-1} |w_\alpha\rangle\}$ gives the equivalence operator

$$\hat{1}' = \sum_\alpha^N \frac{|w_\alpha\rangle}{|w_\alpha|} \frac{\langle w_\alpha|}{|w_\alpha|} \tag{D.44}$$

for the vector space of $\{|u_n\rangle\}$. The density of states $\hat{n}'(\varepsilon)$ for $\hat{H}'$ is defined by

$$\hat{n}'(\varepsilon) \equiv \delta(\varepsilon - \hat{H}'), \tag{D.45}$$

as the operator. Equation (D.44) gives the projected DOS for the vector $|u\rangle$ as

$$\begin{aligned} n'_{11}(\varepsilon) &\equiv \langle u|\delta(\varepsilon - \hat{H}')|u\rangle \\ &= \sum_{\alpha}^{N} \frac{\langle u|w_\alpha\rangle}{|w_\alpha|}\delta(\varepsilon-\varepsilon_\alpha)\frac{\langle w_\alpha|u\rangle}{|w_\alpha|} \\ &= \sum_{\alpha}^{N} \frac{1}{|w_\alpha|^2}\delta(\varepsilon-\varepsilon_\alpha). \end{aligned} \tag{D.46}$$

Equation (D.44) also gives the projected Green function as

$$\begin{aligned} G'_{11}(\varepsilon+i0) &\equiv \langle u|\frac{1}{\varepsilon+i0-\hat{H}'}|u\rangle \\ &= \sum_{\alpha}^{N} \frac{\langle u|w_\alpha\rangle}{|w_\alpha|}\frac{1}{\varepsilon+i0-\varepsilon_\alpha}\frac{\langle w_\alpha|u\rangle}{|w_\alpha|} \\ &= \sum_{\alpha}^{N} \frac{1}{|w_\alpha|^2}\frac{1}{\varepsilon+i0-\varepsilon_\alpha} \end{aligned} \tag{D.47}$$

From Eqs. (D.46) and (D.47), we can see

$$n'_{11}(\varepsilon) = \frac{-1}{\pi}\mathrm{Im}\left[G'_{11}(\varepsilon+i0)\right]. \tag{D.48}$$

Finally, the projected Green function $G'_{11}(\varepsilon+i0)$ is given as a continued fraction; Let us define $D_n(\varepsilon)$ as the determinant of the partial matrix of the tridiagonal matrix in Eq. (D.31) in the sense that it does *not* contain the first n rows and columns. For example,

$$D_0(\varepsilon) \equiv \begin{vmatrix} \varepsilon-a_1 & -b_1 & & & & & \\ -b_1 & \varepsilon-a_2 & -b_2 & & & & \\ & -b_2 & \varepsilon-a_3 & -b_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -b_{N-2} & \varepsilon-a_{N-1} & -b_{N-1} & \\ & & & & -b_{N-1} & \varepsilon-a_N & \end{vmatrix}, \tag{D.49}$$

$$D_1(\varepsilon) \equiv \begin{vmatrix} \varepsilon-a_2 & -b_2 & & & & \\ -b_2 & \varepsilon-a_3 & -b_3 & & & \\ & \cdots & \cdots & \cdots & & \\ & & -b_{N-2} & \varepsilon-a_{N-1} & -b_{N-1} & \\ & & & -b_{N-1} & \varepsilon-a_N & \end{vmatrix}. \tag{D.50}$$

The Laplace expansion gives the recurrence relation of

$$D_n(\varepsilon) = (\varepsilon-a_{n+1})D_{n+1}(\varepsilon) - b_{n+1}^2 D_{n+1}(\varepsilon). \tag{D.51}$$

Since the inverse of a matrix A is given, with the cofactor $\tilde{A}_{ij}$, as

$$(A^{-1})_{ij} = \frac{1}{\det A}\tilde{A}_{ij}, \tag{D.52}$$

the projected Green function

$$G'_{11}(\varepsilon) \equiv \left[\begin{pmatrix} \varepsilon - a_1 & -b_1 & & & & \\ -b_1 & \varepsilon - a_2 & -b_2 & & & \\ & -b_2 & \varepsilon - a_3 & -b_3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & -b_{N-2} & \varepsilon - a_{N-1} & -b_{N-1} \\ & & & & -b_{N-1} & \varepsilon - a_N \end{pmatrix}^{-1} \right]_{11}, \quad \text{(D.53)}$$

is calculated as

$$G'_{11}(\varepsilon) = \frac{D_1(\varepsilon)}{D_0(\varepsilon)}. \quad \text{(D.54)}$$

With Eq.(D.51), we obtain

$$\begin{aligned} G'_{11}(\varepsilon) &= \frac{D_1(\varepsilon)}{D_0(\varepsilon)} \\ &= \frac{D_1(\varepsilon)}{(\varepsilon - a_1)D_1(\varepsilon) - b_1^2 D_2(\varepsilon)} \\ &= \frac{1}{(\varepsilon - a_1) - b_1^2 \dfrac{D_2(\varepsilon)}{D_1(\varepsilon)}}. \end{aligned} \quad \text{(D.55)}$$

When we use Eq.(D.51) successively, the projected Green function is obtained in the continuum fraction form of

$$\begin{aligned} G'_{11}(\varepsilon) &\equiv \langle u | \frac{1}{\varepsilon - \hat{H}'} | u \rangle \\ &= \cfrac{1}{\varepsilon - a_1 - \cfrac{b_1^2}{\varepsilon - a_2 - \cfrac{b_2^2}{\cdots}}} \end{aligned} \quad \text{(D.56)}$$

Equations (D.56) and (D.48) give the projected DOS $n'_{11}(\varepsilon)$ without calculating eigen values nor eigen vectors.

In practical calculations, the projected Green function for H, not H', is given as

$$\langle u | \frac{1}{\varepsilon + i0 - H} | u \rangle \approx \cfrac{1}{\varepsilon - a_1 - \cfrac{b_1^2}{\varepsilon - a_2 - \cfrac{b_2^2}{\cdots . \cfrac{}{\varepsilon - a_N - \cfrac{b_N^2}{T_N(\varepsilon)}}}}} \quad \text{(D.57)}$$

with the function $T_N(\varepsilon)$ called 'terminator'. Explicit function forms of $T_N(\varepsilon)$ are given in textbooks.

D.4 Verlet algorithm in molecular dynamics

For the numerical integration of the Newton equation in the molecular dynamics simulations, we use the velocity-Verlet algorithm, the velocity version of the Verlet algorithm [144]. This section is devoted to a brief description of the velocity-Verlet algorithm for the micro-canonical and canonical ensembles.

Algorithm for micro-canonical ensemble

First we introduce the velocity-Verlet algorithm for the micro-canonical ensemble. We consider a system with N_A classical particles, where the total energy and the Newton equation are given by

$$E = \sum_I \frac{M_I}{2} \dot{\boldsymbol{R}}_{\boldsymbol{I}}^{\,2} + U(\{\boldsymbol{R}_{\boldsymbol{I}}\}) \tag{D.58}$$

$$\ddot{\boldsymbol{R}}_I = \boldsymbol{A}_I \equiv -\frac{1}{M_I} \frac{\partial U}{\partial \boldsymbol{R}_{\boldsymbol{I}}}, \tag{D.59}$$

respectively. The corresponding velocity-Verlet algorithm is as follows

$$\dot{\boldsymbol{R}}_I(t) = \dot{\boldsymbol{R}}_I(t-h) + \frac{h}{2}\left\{\boldsymbol{A}_I(t) + \boldsymbol{A}_I(t-h)\right\} + O(h^3) \tag{D.60}$$

$$\boldsymbol{R}_I(t+h) = \boldsymbol{R}_I(t) + h\dot{\boldsymbol{R}}_I(t) + \frac{h^2}{2}\boldsymbol{A}_I(t) + O(h^3) \tag{D.61}$$

These are equivalent to the second order Taylor expansion, which is seen using the relation $\boldsymbol{A}_I(t) = \boldsymbol{A}_I(t-h) + h\dot{\boldsymbol{A}}_I(t-h) + O(h^2)$. The flow chart of calculations are given by

$$\Rightarrow \boldsymbol{R}_I(t) \Rightarrow \boldsymbol{A}_I(t) \Rightarrow \dot{\boldsymbol{R}}_I(t) \Rightarrow \boldsymbol{R}_I(t+h) \Rightarrow \tag{D.62}$$

The error of the total-energy conservation at the n-th MD time step ($t = nh$) is estimated as

$$\delta E_n \equiv E(nh) - E((n-1)h) \propto h^3 \tag{D.63}$$

for one MD step. For a finite time evolution with a finite time-interval τ, the number of MD steps is $\nu \equiv (\tau/h)$. So the errors at all MD steps are summed up to be

$$E(\tau = \nu h) - E(0) = \sum_n^{\nu} \delta E_n \propto \nu h^3 = \left(\frac{\tau}{h}\right) h^3 = \tau h^2 \tag{D.64}$$

Algorithm for canonical ensemble

Now we turn to explain the finite-temperature dynamics within the Nosé 'thermostat' method [123, 124]. We do not derive the formulation and just show the resultant Newton equation with the temperature T;

$$\ddot{\boldsymbol{R}}_I = \boldsymbol{A}_I - \dot{\eta}\dot{\boldsymbol{R}}_{\boldsymbol{I}} \tag{D.65}$$

$$\ddot{\eta} = \frac{1}{Q}\left[\sum_I M_I \dot{\boldsymbol{R}}_{\boldsymbol{I}}^{\,2} - 3N_A k_B T\right]. \tag{D.66}$$

The conserved energetic quantity is given by

$$H^* = \sum_I \frac{M_I}{2}\dot{\boldsymbol{R}}_{\boldsymbol{I}}{}^2 + U(\{\boldsymbol{R}_{\boldsymbol{I}}\}) + \frac{Q}{2}\dot{\eta}^2 + 3N_A k_B T\eta, \tag{D.67}$$

where the third and fourth terms can be interpreted as the kinetic and potential energies of the 'thermostat' freedom η. The parameter Q corresponds to the mass of the 'thermostat'. One can prove that the time average of the present dynamics gives the ensemble average in the canonical ensemble [123, 124]. As a practical viewpoint, if we choose a proper value for Q, the 'thermostat' works well and the kinetic energy is expected to be almost constant:

$$\sum_I \frac{M_I}{2}\dot{\boldsymbol{R}}_{\boldsymbol{I}}{}^2 \approx \frac{3}{2}N_A k_B T. \tag{D.68}$$

The corresponding velocity-Verlet algorithm is as follows;

$$\begin{aligned}
\dot{\eta}(t) &= \dot{\eta}(t-h) \\
&\quad + \frac{h}{2Q}\left[\sum_I M_I \dot{\boldsymbol{R}}_I(t)^2 + \sum_I M_I \dot{\boldsymbol{R}}_I(t-h)^2 - 6N_A k_B T\right] && \text{(D.69)} \\
\dot{\boldsymbol{R}}_I(t) &= \dot{\boldsymbol{R}}_I(t-h) \\
&\quad + \frac{h}{2}\left[\boldsymbol{A}_I(t) + \boldsymbol{A}_I(t-h) - \dot{\eta}(t)\dot{\boldsymbol{R}}_I(t) - \dot{\eta}(t-h)\dot{\boldsymbol{R}}_I(t-h)\right] && \text{(D.70)} \\
\eta(t+h) &= \eta(t) + h\dot{\eta}(t) + \frac{h^2}{2Q}\left[\sum_I M_I \dot{\boldsymbol{R}}_I(t)^2 - 3N_A k_B T\right] && \text{(D.71)} \\
\boldsymbol{R}_I(t+h) &= \boldsymbol{R}_I(t) + h\dot{\boldsymbol{R}}_I(t) + \frac{h^2}{2}\left[\boldsymbol{A}_I(t) - \dot{\eta}(t)\dot{\boldsymbol{R}}_I(t)\right]. && \text{(D.72)}
\end{aligned}$$

Equations (D.69) and (D.70) are implicit formula and are rewritten as

$$\begin{aligned}
0 &= \dot{\eta}(t) - \dot{\eta}(t-h) - \frac{h}{2Q}\left[K + \frac{K^*}{(1+\frac{h}{2}\dot{\eta}(t))^2} - 6N_A k_B T\right] && \text{(D.73)} \\
\dot{\boldsymbol{R}}_I(t) &= \frac{1}{1+\frac{h}{2}\dot{\eta}(t)}\boldsymbol{W}_I(t), && \text{(D.74)}
\end{aligned}$$

respectively, where

$$\begin{aligned}
K &\equiv \sum_I M_I \dot{\boldsymbol{R}}_I{}^2(t-h) && \text{(D.75)} \\
K^* &\equiv \sum_I M_I \boldsymbol{W}_I{}^2(t) && \text{(D.76)} \\
\boldsymbol{W}_I(t) &\equiv \left\{1 - \frac{h}{2}\dot{\eta}(t-h)\right\}\dot{\boldsymbol{R}}_I(t-h) + \frac{h}{2}\left\{\boldsymbol{A}_I(t) + \boldsymbol{A}_I(t-h)\right\}. && \text{(D.77)}
\end{aligned}$$

Equation (D.73) is the equation for $x \equiv \dot{\eta}(t)$ and is solved iteratively using the Newton-Raphson method;

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})} \tag{D.78}$$

where j is the number of the iterations and

$$f(x) \equiv x - \dot{\eta}(t-h) - \frac{h}{2Q}\left[K + \frac{K^*}{\left(1+\frac{h}{2}x\right)^2} - 6N_A k_B T\right] \tag{D.79}$$

$$f'(x) = 1 + \frac{h^2}{2Q}\frac{K^*}{(1+\frac{h}{2}x)^3}. \tag{D.80}$$

This iterative algorithm needs a proper initial value $x^{(0)}$ and we choose the value to be

$$x^{(0)} = \dot{\eta}(t-h) + \frac{h}{Q}\left[\sum_I M_I \dot{\boldsymbol{R}}_I{}^2(t-h) - 3N_A k_B T\right], \tag{D.81}$$

which is given by Eq.(D.69) under the assumption of

$$\sum_I M_I \dot{\boldsymbol{R}}_I{}^2(t-h) \approx \sum_I M_I \dot{\boldsymbol{R}}_I{}^2(t). \tag{D.82}$$

The algorithm is summarized as follows, where the variables $\{\boldsymbol{R}_I(t), \dot{\boldsymbol{R}}_I(t-h), \boldsymbol{R}_I(t-h), \boldsymbol{A}_I(t), \boldsymbol{A}_I(t-h), \dot{\eta}(t-h), \eta(t-h)\}$ are already obtained:

1. Calculate $\dot{\eta}(t)$ using the Newton-Raphson method with Eq. (D.78) and the initial value of Eq. (D.81).
2. Calculate $\{\dot{\boldsymbol{R}}_I(t)\}$ using Eq. (D.74).
3. Determine $\eta(t+h)$ and $\{\boldsymbol{R}_I(t+h)\}$ using Eqs. (D.71) and (D.72) .

Note that if we set the values of η and $\dot{\eta}$ to be zero at the all time steps, Eqs.(D.70) and (D.72) are reduced to Eqs.(D.60) and (D.61).

Acknowledgements

The author thanks Prof. Takeo Fujiwara for his guidance and encouragement throughout the present thesis. The author also thanks Susumu Yamamoto for general discussions on physics, mathematics, numerical methods, and so on. The calculation by the recursion method was done in the collaboration with Ryu Takayama. The parallelization within the MPI technique was done in the collaboration with Masaaki Geshi. The parallel computations were done by the computer facilities at the Japan Atomic Energy Research Institute. Last but not least, the author thanks all the members and ex-members of the Fujiwara laboratory and other laboratories in the theory division (*Rikigaku Kyoushitsu*) of the Department of Applied Physics at the University of Tokyo.

Bibliography

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964).

[2] W. Kohn and L. S. Sham, Phys. Rev. **140A**, 1133 (1965).

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

[4] International Technology Roadmap for Semiconductors 2002 Update, http://public.itrs.net .

[5] T. Hoshi and T. Fujiwara, preprint (cond-mat 0210366), to appear in J. Phys. Soc. Jpn. **72**, No.10 (2003) .

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

[7] J. F. Janak, Phys. Rev. **B18**, 7165 (1978).

[8] J. O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).

[9] N. D. Mermin, Phys. Rev. **137**, A1441 (1965).

[10] W. Kohn and P. Vashista, in *Theory of the inhomogeneous electron gas*, ed. S. Lundqvist and N. H. March, Plenum, New York (1983).

[11] Y. Yamamoto and T. Fujiwara, Phys. Rev. **B46**, 13596 (1992).

[12] O. K. Anderson, Phys. Rev. **B12**, 3060 (1975).

[13] O. K. Andersen, O. Jepsen, and D. Glötzel, in *Highlights of condensed matter theory*, North Holland (1985).

[14] D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[15] G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. **B26**, 4199 (1982).

[16] D. Vanderbilt, Phys. Rev. **B41**, 7892 (1990).

[17] K. Laasonen, A. Pasquarello, R. Car, C. Lee and D. Vanderbilt, Phys. Rev. **B47**, 10142 (1993).

[18] H. J. Nowak, O. K. Andersen, T. Fujiwara, O. Jepsen and P. Vargas, Phys. Rev. **B44**, 3577 (1991).

[19] A. I. Lichtenstein and M. I. Katsnelson, Phys. Rev. **B57**, 6884 (1998).

[20] A. I. Lichtenstein, M. I. Katsnelson, and G. Kotliar, Phys. Rev. Lett. **87**, 067205 (2001).

[21] W. Kohn, Phys. Rev. **B7**, 4388 (1973).

[22] W. Kohn, Chem. Phys. Lett. **208**, 167 (1993).

[23] G. H. Wannier, Phys. Rev. **52**, 191 (1937).

[24] W. Kohn, Phys. Rev. **115**, 809 (1959).

[25] F. Mauri, G. Galli, and R. Car, Phys. Rev. **B47**, 9973 (1993).

[26] P. Ordejón, D. A. Drabold, M. P. Grumbach, and R. Martin, Phys. Rev. **B48**, 14646 (1993).

[27] T. Hoshi and T. Fujiwara, J. Phys. Soc. Jpn. **69**, 3773 (2000).

[28] C. Edmiston and K. Ruedenberg, Rev. Mod. Phys. **35**, 457 (1963).

[29] S. F. Boys, in *Quantum theory of atoms, molecules and the solid states*, ed. P. Lödin, Academic Press, London, 253 (1966).

[30] C. Edmiston and K. Ruedenberg, in *Quantum theory of atoms, molecules and the solid states*, ed. P. Lödin, Academic Press, London, 263 (1966).

[31] E. Steiner, *The determination and interpretation of molecular wave funtions*, Cambridge University Press (1976).

[32] N. Marzari and D. Vanderbilt, Phys. Rev. **B56**, 12847 (1997).

[33] S. Fujinaga, *Bunshi-kidou hou (Molecular orbital methods)*, Iwanami Shoten, Tokyo [Written in japanese] (1980).

[34] T. Hughbanks and R. Hoffmann, J. Am. Chem. Soc. **105**, 3528 (1983).

[35] R. Dronskowski and P. E. Blöchl, J. Phys. Chem **97**, 8617 (1993).

[36] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).

[37] S. Goedecker, Phys. Rev. **B58**, 3501 (1998).

[38] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists*, fourth ed., Academic Press, San Diego (1995).

[39] S. Roche, Phys. Rev. **B59**, 2284 (1999).

[40] RIKEN REVIEW **29**, (2000).

[41] P. Ordejón, Comp. Mat. Sci. **12**, 157 (1998).

[42] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).

[43] G. Galli, Phys. Stat. Sol. (b) **217**, 231 (2000).

[44] S. Y. Wu and C. S. Jayanthi, Phys. Rep. **358**, 1 (2002).

[45] D. R. Bowler, M. Aoki, C. M. Goringe, A. P. Horsfield and D. G. Pettifor, Modelling Simul. Mater. Sci. Eng. **5**, 199 (1997).

[46] X. P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. **B47**, 10891 (1993).

[47] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).

[48] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).

[49] S. Goedecker and M. Teter, Phys. Rev. **B51**, 9455 (1995).

[50] D. G. Peffifor, Phys. Rev. Lett. **63**, 2480 (1989).

[51] R. Haydock, *The recursive solution of the Schrödinger equaiton'*, in *Solid state physics*, ed. H. Ehrenreich, F. Seitz, D. Turnbull **35**, 215 (1980).

[52] R. Takayama, T. Hoshi, and T. Fujiwara, in preparation .

[53] L. D. Landau and E. M. Lifshitz, *Quantum mechanics (non-relativistic theory)*, 3rd ed., Pergamon Press, Oxford (1976).

[54] N. W. Ashcroft and N. D. Mermin, *Solid state physics*, Saunders College, Philadelphia (1976).

[55] W. Stich, E. K. U. Gross, P. Malzacher, and R. M. Dreizler, Z. Phys. A **309**, 5 (1982).

[56] R. M. Dreizler and E. K. U. Gross, *Density functional theory*, Springer-Verlag, Berlin (1990).

[57] M. Pearson, E. Smargiassi, and P. A. Madden, J. Phys.: Condens. Matter **5**, 3221 (1993).

[58] J. A. Anta, B. J. Jesson, and P. A. Madden, Phys. Rev. **B58**, 6124 (1998).

[59] J. C. Phillips, Rev. Mod. Phys. **42**, 317 (1970).

[60] J. S. Slator and G. F. Koster, Phys. Rev. **B94**, 1498 (1954).

[61] P.Vogl, H. P. Hjalmarson, and J. D. Dow, J. Phys. Chem. Solids **44**, 365 (1983).

[62] W. A. Harrison, *Electronic structure and the properties of solids*, W. H. Freeman and Company, San Fransisco (1980).

[63] I. Stich, R. Car, and M. Parrinello, Phys. Rev. Lett. **63**, 2240 (1989).

[64] C. H. Xu, C. Z. Wang, C. T. Chan, and K. M. Ho, J. Phys. Condens. Matter **4**, 6047 (1992).

[65] D. J. Chadi, Phys. Rev. Lett. **43**, 43 (1979).

[66] A. Ramstad, G. Brocks, and P. J. Kelly, Phys. Rev. **B51**, 14505 (1995).

[67] R. A. Wolkow, Phys. Rev. Lett. **92**, 2636 (1992).

[68] D. J. Chadi, J. Vac. Sci. Technol. **16**, 1290 (1979).

[69] D. J. Chadi and M. L. Cohen, Phys. Status. Solidi (b) **68**, 405 (1975).

[70] P. Kröger and J. Pollmann, Phys. Rev. Lett. **74**, 1155 (1995).

[71] Z. Zhu, N. Shima, and M. Tsukada, Phys. Rev. **B40**, 11868 (1989).

[72] E. Mooser and W. B. Pearson, Acta Crysta. **12**, 1015 (1959).

[73] A. I. Shkrebtii and R. D. Sole, Phys. Rev. Lett. **70**, 2645 (1993).

[74] S. Liu, C. S. Jayanthi, S-Y. Wu, X. Qin, Z. Zhang and M. G. Lagally, Phys. Rev. **B61**, 4421 (2000).

[75] R. W. Tank and C. Arcangeli, Phys, Stat. Sol. (b) **217**, 89 (2000).

[76] O. K. Andersen and T. Saha-Dasgupta, Phys. Rev. **B62**, R16219 (2000).

[77] O. K. Andersen, T. Saha-Dasgupta, R. W. Tank, C. Arcangeli, O. Jepsen and G. Krier, in *Electronic Structure and Physical Properties of Solids. The Uses of the LMTO Method*, Ed. H. Dreysse, Springer-Verlag, Berlin, 3 (2000).

[78] O. K. Andersen, T. Saha-Dasgupta, and S. Ezhov, Bull. Mater. Sci. **26**, 19 (2003).

[79] D. Nguyen-Manh, T. Saha-Dasgupta, and O. K. Andersen, Bull. Mater. Sci. **26**, 27 (2003).

[80] M. Geshi, T. Hoshi, and T. Fujiwara, preprint (cond-mat/0306461) .

[81] O. H. Nielsen and R. M. Martin, Phys. Rev. **B32**, 3792 (1985).

[82] F. H. Stillinger and T. A. Weber, Phys. Rev. **B31**, 5262 (1985).

[83] http://www.mpi-forum.org/ .

[84] http://www.openmp.org/ .

[85] T. Hoshi and T. Fujiwara, Surf. Sci. **493**, 659 (2001).

[86] U. Stephan and D. A. Drabold, Phys. Rev. **B57**, 6391 (1998).

[87] D. Weaire and M. F. Thorpe, Phys. Rev. **B4**, 2508 (1971).

[88] M. F. Thorpe and D. Weaire, Phys. Rev. **B4**, 3518 (1971).

[89] A. A. Griffith, Philos. Trans. R. Soc. London Ser. A **221**, (1920).

[90] H. Liebowitz (ed.), *Fracture*, Academic Press, New York **I-VII**, (1968-1972).

[91] L. B. Freund, *Dynamic fracture mechanics*, Cambridge University Press (1989).

[92] B. Lawn, *Fracture of brittle solids*, 2nd ed., Cambridge University Press (1993).

[93] R. Thomson, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull, Academic Press, New York 1 (1986).

[94] C. John, Phil. Mag. **32**, 1193 (1975).

[95] M. Brede and P. Haasen, Acta Metall. **36**, 2003 (1988).

[96] J. Samuels and S. G. Roberts, Proc. R. Soc. Lond. **A421**, 25 (1989).

[97] K. C. Pandey, Phys. Rev. Lett. **47**, 1913 (1981).

[98] K. C. Pandey, Phys. Rev. Lett. **49**, 223 (1982).

[99] J. E. Northrup and M. Cohen, Phys. Rev. Lett. **49**, 1349 (1982).

[100] F. Ancilotto, W. Andreoni, A. Selloni, R. Car and M Parrinello, Phys. Rev. Lett. **65**, 3148 (1990).

[101] L. D. Landau and E. M. Lifshitz, *Statistical physics*, 3rd ed. Part I, Pergamon Press, Oxford (1980).

[102] J. C. H. Spence, Y. M. Huang, and O. Sankey, Acta Metall. Mater. **41**, 2815 (1993).

[103] R. Pérez and P. Gumbsch, Phys. Rev. Lett. **84**, 5347 (2000).

[104] R. Pérez and P. Gumbsch, Acta Mater. **48**, 4517 (2000).

[105] T. Cramer, A. Wanner, and P. Gumbsch, Phys. Rev. Lett. **85**, 788 (2000).

[106] J. Tersoff, Phys. Rev. **B37**, 6691 (1988).

[107] J. Tersoff, Phys. Rev. **B38**, 9902 (1988).

[108] H. Balamane, T. Halicioglu, and W. A. Tiller, Phys. Rev. **B46**, 2250 (1992).

[109] D. Holland and M. Marder, Phys. Rev. Lett. **80**, 746 (1998).

[110] D. Holland and M. Marder, Adv. Mater. **11**, 793 (1999).

[111] C. C. Fu, M. Weissman, and A. Saúl, Surf. Sci. **B494**, 119 (2001).

[112] J. Ihm, D. Lee, J. D. Joannopoulos, and J. J. Xiong, Phys. Rev. Lett. **51**, 1872 (1983).

[113] K. Inoue, Y. Morikawa, K. Terakura, and M. Nakayama, Phys. Rev. **B49**, 14774 (1994).

[114] K. Hata, Y. Sainoo, and H. Shigekawa, Phys. Rev. Lett. **86**, 3084 (2001).

[115] H. J. W. Zandvliet, Rev. Mod. Phys. **72**, 593 (2000).

[116] D. J. Chadi, Phys. Rev. Lett. **59**, 1691 (1987).

[117] B. S. Swartzentruber, Y. W. Mo, R. Kariotis, M. G. Lagally and M. B. Webb, Phys. Rev. Lett. **65**, 1913 (1990).

[118] A. Oshiyama, Phys. Rev. Lett. **74**, 130 (1995).

[119] O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopoulos, Phys. Rev. Lett. **61**, 1973 (1988).

[120] A. García and J. E. Northrup, Phys. Rev. **B48**, 17350 (1993).

[121] J. Dabrowski, E. Pehlke, and M. Scheffler, Phys. Rev. **B49**, 4790 (1994).

[122] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, and D. Vanderbilt, Phys. Rev. Lett. **64**, 2406 (1990).

[123] S. Nosé, Mol. Phys. **52**, 255 (1984).

[124] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[125] M. Miyata, T. Fujiwara, S. Yamamoto, and T. Hoshi, Phys. Rev. **B60**, R2135 (1999).

[126] L. Hedin, Phys. Rev. **139**, A796 (1965).

[127] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).

[128] W. E. Pickett and C. S. Wang, Phys. Rev. **B30**, 4719 (1984).

[129] M. Rohlfing, P. Krüger, and J. Pollmann, Phys. Rev. **B52**, 1905 (1995).

[130] A. Yamasaki, D. Thesis, Dept. of Appl. Phys, Univ. of Tokyo [Written in japanese] (2002).

[131] C. Kittel, *Introduction to solid state physics*, 7th ed., John Wiley & Sons Inc., New York (1996).

[132] L. D. Landau and E. M. Lifshitz, *Theory of elasticity*, 2nd ed., Pergamon Press, Oxford (1970).

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

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

[135] C. S. G. Cousins, L. Gerward, J. S. Olsen, B. Selsmark and B.J. Sheldon, J. Phys. C **20**, 29 (1987).

[136] A. E. H. Love, *A treatise on the mathematical theory of elasticity*, 4th ed., Dover Publications, New York (1944).

[137] D. Maugis, *Contact, adhesion and rupture of elastic solids*, Springer-Verlag, Berlin (1999).

[138] G. R. Irwin, Fracture, in *Handbuch der Physik*, Springer-Verlag, Berlin **6**, 551 (1958).

[139] N. F. Mott, Engeering **165**, 16 (1948).

[140] L. B. Freund, J. Geophys. Res. **84**, 2199 (1979).

[141] K. B. Broberg, Int. J. Fract. **39**, 1 (1989).

[142] F. Abraham and H. Gao, Phys. Rev. Lett. **84**, 3113 (2000).

[143] C. Lanczos, J .Res. Natl. Bur. Stand **45**, 255 (1950).

[144] L. Verlet, Phys. Rev. **159**, 98 (1960).