On-The-Fly Machine Learning of Quantum Mechanical Forces and its Potential Applications for Large Scale Molecular Dynamics

Li, Zhenwei

Awarding institution:
King's College London

KING'S COLLEGE LONDON

On-The-Fly Machine Learning of Quantum Mechanical Forces and its Potential Applications for Large Scale Molecular Dynamics

Zhenwei Li

A Dissertation for the
degree of Doctor of Philosophy

in the
School of Natural and Mathematical Sciences
Department of Physics

October 2014

Acknowledgements

This work described in this thesis was carried out during my PhD study in King's College London which provides an international study environment, where I gained many enriching experiences both socially and academically.

I would like to acknowledge my supervisor Prof. Alessandro De Vita, who has given kind help in supervision of my research from the beginning of my PhD. I also owe my thanks to Dr James Kermode for his instruction throughout my PhD time as well as his reading of this thesis. The useful discussions with Gábor Csányi, Albert Bartók, Marco Caccin, Mike Payne, and Mike Finnis are acknowledged regarding this research work. I would acknowledge my PhD examiners : Prof. O. Anatole von Lilienfeld in Basel University and Prof. Zheng-Xiao Guo in University College London.

I would like to thank Julia Kilpatrick, Paul Le Long, Nigel Arnot, James French in the department for their support. The conversations with my colleagues (Giovanni Peralta, Massimo Riello, Giovanni Donni, Dominic Botten, Federico Bianchini, etc.) in the department are an enjoyable memory of my time at King's.

Last but not least, I would like to acknowledge funding from the *Rio Tinto Centre for Advanced Mineral Recovery* based at Imperial College London and additional funding from King's College London.

Preface

This thesis entitled 'On-The-Fly Machine Learning of Quantum Mechanical Forces and its Potential Applications for Large Scale Molecular Dynamics' was carried out under the supervision of Prof. Alessandro De Vita and Dr James Kermode during 10/2010 − 06/2014 in King's College London. I contributed to the following publications:

- Multi-scale modelling of materials chemomechanics: brittle fracture of oxides and semiconductors. James Kermode, Giovanni Peralta, Zhenwei Li, and Alessandro De Vita, Procedia Material Science 3, 2681-1686 (2014).
- Inversion in crack speed as a function of temperature during crossover from activated to catastrophic fracture. Giovanni Peralta, James Kermode, Silvia Cereda, Zhenwei Li, Albert Bartók, Gabor Csányi, and Alessandro De Vita. In Preparation (2014). Presented in Chapter 3.
- Molecular Dynamics with On-The-Fly Machine Learning of Quantum Mechanical Forces. Zhenwei Li, James Kermode and Alessandro De Vita, submitted (2014). Some related materials are presented in Chapters 5, 6 and 7.

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for any other degree or qualification in this, or any other university. This dissertation is the result of my own work and all the outcomes of collaboration are referenced or acknowledged.

KING'S COLLEGE LONDON

Abstract

School of Natural and Mathematical Sciences

Department of Physics

Doctor of Philosophy

Zhenwei Li

Material simulation using molecular dynamics (MD) at the quantum mechanical (QM) accuracy level has gained great interest in the community. However, the bottleneck arising from the $O(N^3)$ scaling of QM calculation has enormously limited its investigation scope. As an approach to address this issue, in this thesis, I proposed a machine-learning (ML) MD scheme based on Bayesian inference from CPU-intensive QM force database. In this scheme, QM calculations are only performed when necessary and used to augment the ML database for more challenging prediction case. The scheme is generally transferable to new chemical situations and database completeness is never required. To achieve the maximal ML efficiency, I use a symmetrically reduced internal-vector representation for the atomic configurations.

Significant speed-up factor is achieved under controllable accuracy tolerance in the MD simulation on test case of Silicon at different temperatures. As the database grows in configuration space, the extrapolative capability systematically increases and QM calculations are finally not needed for simple chemical processes. In the on-the-fly ML force calculation scheme, sorting/selecting out the closest data configurations is used to enhance the overall efficiency to scale as $\sim O(N)$. The potential application of this methodology for large-scale simulation (e.g. fracture, amorphous, defect), where chemical accuracy and computational efficiency are required at the same time, can be anticipated.

In the context of fracture simulations, a typical multi-scale system, interesting events happen near the crack tips beyond the description of classical potentials. The simulation results by machine-learning potential derived from a fixed database with no enforced QM accuracy inspire a theoretical model which is further used to investigate the atomic bond breaking process during fracture propagation as well as its relation with the initialised vibration modes, crack speed, and bonding structure.

Contents

List of Figures

Abbreviations

ML	**M**achine **L**earning
QM	**Q**uantum **M**echanics
MD	**M**olecular **D**ynamics
MM	**M**olecular **M**echanics
HF	**H**ellmann **F**eynman
HK	**H**ohenberg **K**ohn
KS	**K**ohn **S**ham
GAP	**G**aussian **A**pproximation **P**otential
NN	**N**eural **N**etworks
BO	**B**orn **O**ppenhemer
LJ	**L**ennard **J**ones
SW	**S**tillinger **W**eber
PBC	**P**eriodic **B**oundary **C**onditions
PES	**P**otential **E**nergy **S**urface
DFT	**D**ensity **F**unctional **T**heory
DFTB	**D**ensity **F**unctional **T**ight **B**inding
LOTF	**L**earn **O**n **T**he **F**ly
MLOTF	**M**achine **L**earning **O**n **T**he **F**ly
FPMD	**F**irst **P**rinciples **M**olecular **D**ynamics

Chapter 1

Introduction

Quantum mechanics provides an accurate description of material properties from the electronic level. Density Functional Theory (DFT), has become the standard approach for performing quantum mechanical simulation of materials from first-principles and has rendered a huge number of publications since its establishment in the 1960s [1, 2]. In the material simulation community, there is an increasing need to explore the atomistic processes, using first-principles molecular dynamics (FPMD). As the scope of such investigations extends, both on temporal and spacial scales, the $O(N^3)$ scaling of FPMD typically becomes a limitation, and it is due to this that, only systems of a few hundred of atoms and/or timescales up to pico seconds can be addressed [3, 4].

Molecular Dynamics (MD) using classical potentials has been used for a long period of time. This method is still appealing nowadays for its efficiency and capability to describe the atomic systems up to millions of atoms in a computer environment. Classical potentials are usually derived by encoding a physical description of the atomic interactions into an analytical functional form whose parameters are fitted with respect to experimental properties, such as elastic constants, bulk modulus and lattice constants. Using these empirical parameters, they typically have applications limited to the domain of problems related to where the potentials were fitted to benchmark properties. Following the work by Ercolessi and Adams in 1994 [5], there was a trend to use force-matching to parameterise potentials. These potentials were generated by adjusting their parameters to match the classical forces with target quantum mechanical (QM) forces derived from

first-principles calculations. Even though parameters were involved, they did produce good quality potentials for metals, semiconductors and oxides [6–8].

In recent years, there have approaches proposed by applying 'machine learning' (ML) techniques to fit the first-principles potential energy surface (PES). These potentials work through functional inference from a QM database 'once-and-for-all'. Among them, the Gaussian Approximation Potentials (GAP) adopt the Gaussian Process function inference [9], and Neural-Networks (NN) potentials [10] use the generalised neural networks techniques. A number of potentials have been generated under these machine-learning schemes with accuracy comparable to the DFT level without performing the self-consistent electronic calculations. These potentials however, in many aspects resemble classical potentials after training with QM data. The atomic forces are calculated by analytical differentiation of the energy. Transferability of these potentials largely relies on the set of chemical environments that can be represented in the database.

Multi-scale problems are challenging essentially because of the long-range stress field, for which a vast number of atoms have to be incorporated into a simulation, while for the chemically active region, quantum accuracy is mandatory for the correct description of bond breaking/forming events. For these simulations, it was proposed to hybrid two different kinds of descriptions, Quantum Mechanics (QM) and Molecular Mechanics (MM). For the mismatching of two such distinct descriptions, different strategies of mixing were developed, such as mechanical mixing or energy mixing [11]. Time embedding schemes, such as the 'Learn-on-the-fly' (LOTF) MD accelerate the MD simulations by adjusting the classical parameters with informative QM calculations only required once every n steps ($n \sim 10$ in Silicon fracture simulation) with force accuracy further enhanced by implementation of predictor-corrector algorithm [12, 13].

The accuracy for a large scale system is still limited when addressing completely new chemical environments either due to the issues with the transferability or the completeness of the database. In this thesis, I will describe a proposed approach that aimed to abstract the maximum transferable knowledge from a QM-force database comprised of computationally expensively data in order to run large-scale MD simulations where highly accurate atomic forces are required. This approach works by performing function inference on the QM force vectors in a straightforward way with no invocation of the energy expressions (either atomic energy or total energy). The chemical environments

that are novel to the database are computed with QM routines when and only when necessary and are used to augment the existing database to enhance the ML prediction capability. This scheme is implemented in such a way that the ML force prediction is always carried out 'on-the-fly'. The QM database built from different MD runs can be used with transferability in force prediction for relevant systems while the prediction accuracy systematically increases as more QM force information is added into the database.

To apply the machine learning of QM force while achieving the maximum ML efficiency, an internal-vector representation for the local chemical environments which are formed by the geometry of the interacting neighbouring atoms was proposed and constructed taking into account the symmetries associated with the atomic force vector quantity. This representation also makes it practical to incorporate information from additional vectors e.g. the commonly used classical or empirical force vectors, leading to systematic improvements with respect to the QM benchmark.

The results yielded by MLOTF in this work demonstrate a systematic increase in efficiency and accuracy as the database grows during MD simulations. Large speed-up factors (e.g. 30 times in the case of Silicon MD at 1000K) compared with the full QM calculations were achieved with controllable accuracy. The force prediction capability is also largely improved upon the previous non-learning LOTF MD. As *data* configurations closer to the prediction configurations are available, force prediction using a smaller subset of the database can be used to make predictions with desirable accuracy. Sorting/selecting the most relevant configurations enables dynamical machine learning and prediction even for huge database (the order of magnitude of millions of atomic configurations). The MLOTF computational cost has a scaling factor close to $O(N)$, which makes it a promising application for large-scale MD simulations, as is to be presented in Section 6.6.

The structure of this thesis is as follows: Background for the Quantum Mechanics (QM) , Density Functional Theory (DFT) and Molecular Dynamics (MD) simulations will be presented in **Chapter 2**. The methodologies beyond classical potentials, including the QM/MM embedding and 'learn-on-the-fly' (LOTF) MD will be described in the context of multi-scale simulations, e.g. in fracture simulations. In **Chapter 3**, fracture

simulation will be explained and the work based on embedding GAP and the Stillinger-Weber potential to investigate propagation speed will be introduced. Furthermore, a theoretical model will also be described to further probe the mechanism associated with bond breaking in brittle fracture. In **Chapter 4**, background for the Gaussian Process function inference will be discussed as well as an overview of the machine learning techniques. In the later part of this chapter, machine-learning (ML) potentials such as, the Gaussian Approximation Potentials, Neural-Network potentials as well as an ML scheme for calculating the atomisation energy in a molecular compound will be introduced as background for the work described in the following chapters. In **Chapter 5**, an approach to machine learning of QM forces will be proposed and constructed. This chapter starts from the symmetrically-reduced representation for atomic environments before moving to implementation of Gaussian Process inference into atomistic force prediction. The feature of performing the force prediction will also be explained targeting practical applications in large scale MD simulations. In **Chapter 6**, the methodology developed for force calculations will be systematically tested using a static database, including the application into phonon calculation. In **Chapter 7**, the force calculation will be applied into large-scale MD in an 'on-the-fly' manner. As an improvement upon the non-learning LOTF calculation, the accuracy and efficiency in our new force calculation scheme has enhanced learning capability with an dynamically updating database. In **Chapter 8**, the application of the methodology will be extended into more complex binary-system taking SiC and SiO_2 as examples. Preliminary results for these systems will be presented.

Chapter 2

Background- I

2.1 Quantum Description

Quantum Mechanics (QM) opens possibilities to investigate the microscopic physics with unified description of both the particle and wave natures of matter. In the following part, I give a review of the QM description and the theorems that enable the approximately accurate QM simulation of multi-body systems.

In QM, particle dynamics is expressed by the Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\psi(\vec{r},t) = \hat{H}\psi(\vec{r},t) \tag{2.1}$$

where $\psi(\vec{r},t)$ is the electronic wavefunction, t the time, $\hbar$ equals the Planck constant divided by 2π, and $\hat{H}$ is the Hamiltonian operator which can be written as the kinetic and potential parts:

$$\hat{H} = \hat{T} + \hat{V}(\vec{r},t) \tag{2.2}$$

For time-invariant potential $\hat{V}(\vec{r})$, the ground state Schrödinger equation is:

$$\hat{H}\psi(\vec{r}) = \epsilon\psi(\vec{r}) \tag{2.3}$$

where ϵ indicates the eigen energy value.

2.1.1 Adiabatic Approximation

For atomic system, the Hamiltonian $\hat{H}$ is comprised of both interactions among electrons $\{\vec{r}_i\}$ and ions $\{\vec{R}_I\}$:

$$\hat{H}_{BO} = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 + \frac{e^2}{2}\sum_{i<j}\frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{i,I}\frac{Z_I \cdot e^2}{|\vec{r}_i - \vec{R}_I|} - \frac{\hbar^2}{2M_I}\sum_{I}\nabla_I^2 + \frac{e^2}{2}\sum_{I<J}\frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|} \tag{2.4}$$

where the first and forth term indicates the kinetic operators from electrons and ions respectively, the second term corresponds to the Coulomb interaction between each pair of electrons and the third term gives the Coulomb interaction between electrons and ions, and fifth term the Coulomb interaction between ions. In the equation, the index i runs over the N electrons, while Z_I and Z_J correspond to the atomic number of ions I and J. m and M are the mass of the electrons and ions, respectively. $\vec{r}$ and $\vec{R}$ indicate the electronic and ionic coordinates, respectively.

The time-invariant Schrödinger equation is,

$$\hat{H}_{BO}|\Psi_{BO}\rangle = E|\Psi_{BO}\rangle \tag{2.5}$$

For more complex system than Hydrogen, the exact solution for the above multi-body equation becomes prohibitive to be attained. To address this issue, Born and Oppenheimer proposed in 1927 that, the Hamiltonian derived from ionic and electronic coordinates can be separated, based on the fact that their relaxation time scales usually differ by several orders of magnitude, or to say, $\tau_e << \tau_R$ (see Ref.[14]). Under the Born-Oppenheimer (BO) approximation, the electronic states can be explicitly solved with a Hamiltonian incorporating the potential energy determined by the stationary ionic coordinate. The dynamics of the ions can be constructed according to the ground electronic states, as in first-principles molecular dynamics to be discussed in next Chapter.

According to the BO approximation, wavefunctions are written as:

$$\Psi_{BO}(\{\vec{r}_i\},\{\vec{R}_I\}) = \Psi(\{\vec{r}_i\},\{\vec{R}_I\})\Theta(\{\vec{R}_I\}). \tag{2.6}$$

where $\Psi(\{\vec{r}_i\},\{\vec{R}_I\})$ indicate the electronic wave-function given the ionic coordinates $\{\vec{R}_I\}$ and $\Theta(\vec{R}_I)$ the wave-function for the ions with electrons always relaxed to the ground state.

Under this approximation, a potential energy surface (PES) can be defined as the energy (both from electron and ion Coulomb interactions) landscape with respect to the different ionic configurations, and thus this PES reveals the stability of the given structure at ambient conditions. The approximation is, however, not valid in the cases where coupling between electrons and ions becomes significant [15–17].

2.1.2 Hartree Fock Scenario

In the Hartree scheme, the N-electron wave-functions is written as product of N single electronic wavefunctions :

$$\Psi_H(r_1,\cdots,r_N) = \psi_1(\vec{r}_1)\psi_2(\vec{r}_2)\cdots\psi_N(\vec{r}_N) \tag{2.7}$$

In this case, the Schrödinger equation becomes:

$$\left[\hat{T}_i + V(\vec{r}_i) + \frac{e^2}{2}\sum_{j\neq i}\int \frac{\rho_j}{|\vec{r}_i-\vec{r}_j|}dr_j^3\right]\psi_i(\vec{r}_i) = \epsilon_i\psi_i(\vec{r}_i) \tag{2.8}$$

where single-electron density $\rho_j \equiv \langle\psi_j(\vec{r}_j)|\psi_j(\vec{r}_j)\rangle$ is the electronic density of j-th electron, and $\hat{T}_i = -\frac{\hbar^2}{2m}\nabla_i^2$ the kinetic operator for i-th electron, $V(\vec{r}_i)$ indicates the external potential upon the i-th electron and the third-term gives the electrostatic potential due to the rest of the electrons, which is known as the Hartree energy.

Instead of using the eigenstates as multiplication of single eigenstates, Fock and Slater proposed to write the partial eigen function (Ψ_{HF}) as an anti-symmetric determinant to describe the N-electron system within the Pauli's principles, as in the following equation:

$$\Psi_{HF} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_N(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) & \cdots & \psi_N(\mathbf{x}_2) \\ \dots & \dots & & \dots \\ \psi_1(\mathbf{x}_N) & \psi_2(\mathbf{x}_N) & \cdots & \psi_N(\mathbf{x}_N) \end{vmatrix} = \frac{1}{\sqrt{N!}}det[\psi_1\psi_2\cdots\psi_N]$$

For spin calculations, each row and column in the Slater form of the Ψ_{HF} (Eq. 2.1.2) are expanded to describe the spin regenerated states.

The Hartree-Fock equation for single electron is written as:

$$\left[\hat{T} + V(\vec{r}) - \frac{e^2}{2}\int dr'^3 \frac{\rho(\vec{r}') - \rho^{HF}(\vec{r},\vec{r}')}{|\vec{r}-\vec{r}'|}\right]\psi(\vec{r}) = \epsilon\psi(\vec{r}) \tag{2.9}$$

where $\hat{T}$ is the kinetic energy operator. The electron density due to the exchange of electrons in the HF Hamiltonian is indicated by the term, $\rho^{HF}(\vec{r},\vec{r}')$.

The single-electron Hamiltonian then becomes:

$$\hat{H}_{HF} = \hat{T} + V(\vec{r}) - \frac{e^2}{2}\int dr'^3 \frac{\rho(\vec{r}') - \rho^{HF}(\vec{r},\vec{r}')}{|\vec{r}-\vec{r}'|} \tag{2.10}$$

Under the Hartree-Fock approximation, the total energy is:

$$E_{HF} = \left\langle \Psi_{HF}|\hat{H}_{HF}|\Psi_{HF}\right\rangle = \sum_i H_i + \frac{1}{2}\sum_{i,j=1}^{N}(J_{ij} - K_{ij}) \tag{2.11}$$

where $H_i = \left\langle \psi_i|\hat{T} + V(\vec{r})|\psi_i\right\rangle$ and, J_{ij} called the Coulomb integral and K_{ij} the correlation integral.

The Hartree-Fock approximation provides an accurate description of the N-electron quantum system, however it is computationally demanding. In the Hartree-Fock equation of Eq.2.9, the weak correlation energy among electrons is not incorporated. In the following, I introduce the Density Functional theory (DFT) which laid foundations for most modern first-principles calculations.

2.1.3 Density Functional Theory

Density Functional Theory (DFT) has become a standard tool for electronic calculations in quantum chemistry and material science. In 1964, Hohenberg and Kohn [1] proposed to use the electron density $\rho(\vec{r})$ as single basic quantity for considering the $N-$electron system located in the external potential V_{ext} [18]. This was the later known as the Hohenberg-Kohn theorem, which I will introduce as follows:

Hohenberg-Kohn (HK) theorem. Let us start from writing the Hamiltonian as $\hat{H} = \hat{T} + \hat{V}_{ee} + V_{ext}$, where $\hat{T} = -\frac{\hbar^2}{2m}\sum_i \nabla_i^2$ is the kinetic operator, $\hat{V}_{ee}$ indicates the interaction potential yielded by the N-electrons, and V_{ext} the external potential, including but not only the ionic potential. The Hamiltonian above is non-disputive, since the N- electrons and the external potential completely determine the properties of the system, and therefore the Hamiltonian. The total energy is expressed as:

$$E[\rho(\vec{r}), V_{ext}(\vec{r})] = \left\langle \Psi \right| \hat{T} + \hat{V}_{ee} \left| \Psi \right\rangle + \int dr^3 V_{ext}(\vec{r})\rho(\vec{r}) \tag{2.12}$$

where $\rho(\vec{r})$ is the electron density corresponding to the squared modulus of the wave-fuctions. The HK theorem states that the external potential $V_{ext}(\vec{r})$ is determined, within an additive constant, by electron density $\rho(\vec{r})$ [1]. Directly from the theorem, density $\rho(\vec{r})$ is therefore unique feature of the N-electron system.

A simple proof is that, suppose there are two different external potentials V_{ext} and V'_{ext} that correspond to the same electron density $\rho(\vec{r})$. According to the variational principle, it is known that only the Eigen wave-function minimises the associated energy. Two ground-state energies E_0 and E'_0 correspond to the external potentials V_{ext} and V'_{ext}, respectively.

$$\begin{cases} E_0 = \left\langle \Psi | \hat{H} | \Psi \right\rangle < \left\langle \Psi' | \hat{H} | \Psi' \right\rangle = E'_0 + \int dr^3 (V_{ext} - V'_{ext})\rho(\vec{r}) \\ E'_0 = \left\langle \Psi' | \hat{H}' | \Psi' \right\rangle < \left\langle \Psi | \hat{H}' | \Psi \right\rangle = E_0 - \int dr^3 (V_{ext} - V'_{ext})\rho(\vec{r}) \end{cases}$$

Adding up two of the inequalities by each side, we obtain $E'_0 + E_0 < E'_0 + E_0$. This contradictory result suggests that, at ground state the N-electron Hamiltonians must be unique functional of electron density ρ.

Based on this theorem, the ground state properties, such as wavefunctions and energies are all uniquely determined by the electron density $\rho(\vec{r})$. The total energy in Eq.2.12 at ground state is thus written as:

$$E[\rho] = F[\rho] + \int dr^3 V_{ext}(\vec{r})\rho(\vec{r}) \tag{2.13}$$

with $F[\rho]$ corresponding to $F[\rho] = \left\langle \Psi \right| \hat{T} + \hat{V}_{ee} \left| \Psi \right\rangle$ and indicating the sum of the kinetic energy and electron-electron interaction energy for a given electron density distribution

of ρ. This $F[\rho]$ is a universal functional in the sense that it is only determined by the electron density and independent of any external potential.

V-representability. Now we know that system properties can be described using density as a basic quantity. To obtain ground-state ρ, the minimisation of energy function, however, has to be performed on density space satisfying two requirements: (1) N-representable (2) V-representable.

The total energy $E[\tilde{\rho}]$ for the N-electron system with electron density of $\tilde{\rho}$ is:

$$E[\tilde{\rho}] =< \tilde{\Psi}|\hat{T} + V_{\text{ee}} + V_{\text{ext}}|\tilde{\Psi} >= \underbrace{< \tilde{\Psi}|\hat{T} + V_{\text{ee}}|\tilde{\Psi} >}_{F[\tilde{\rho}]} + \underbrace{< \tilde{\Psi}|V_{ext}|\tilde{\Psi} >}_{\int \tilde{\rho}(\vec{r})V_{ext}(\vec{r})dr^3}$$

and $E[\tilde{\rho}] > E[\rho]$ for all the N-representable $\tilde{\rho}$, where ρ are the ground-state electron density.

An N-representable density means that the electron density $\tilde{\rho}$ can be composed by anti-symmetric wavefunctions $\tilde{\Psi}(\vec{r})$. Further minimisation upon the density $\tilde{\rho}$

$$E[\rho] = \min_{\tilde{\rho}\Rightarrow\rho} \left\{ F[\tilde{\rho}] + \int V_{ext}(\vec{r})\tilde{\rho}(\vec{r})dr^3 \right\} \tag{2.14}$$

the ground state energy $E[\rho]$ can be minimised with respect to $\tilde{\rho}$ in the domain of V-representable [18]. A density is V-representable if it is the density associated with the anti-symmetric ground-state wavefunction of a Hamiltonian with *some* external potential $V_{ext}(\vec{r})$. Note, the V-representability of electron density ρ is important for the validation of the minimisation procedure in Eq.2.14.

Kohn-Sham Equation. In 1965, Kohn and Sham [2] found that electronic states in the N-electron system can further be approximated with a single electron that is located in an effective potential V_{eff}. This constructs the famous Kohn-Sham equation as in Eq.2.15. The effective potential incorporates the interaction from all other electrons and external potential while the residual interactions from the single-electron approximation are put into the exchange-correlation term. Therefore, the KS equation is accurate in itself from the quantum mechanical consideration.

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}} \right] \psi_j(\vec{r}) = \epsilon_j \psi_j(\vec{r}) \tag{2.15}$$

The effective potential for the approximated single electron states is:

$$V_{\text{eff}} = V(\vec{r}) + V_{ee} + V_{xc}[\rho(\vec{r})] \tag{2.16}$$

In the above, the $V_{ee} = e^2 \int dr'^3 \frac{\rho(\vec{r'})}{|\vec{r}-\vec{r'}|}$ corresponds to the Hartree-Fock interaction energy, V is the potential fields on electrons and V_{xc} the exchange-correlational potential.

From the KS equation, total energy:

$$E = \sum_j^N \epsilon_j - \frac{e^2}{2} \int dr^3 dr'^3 \frac{\rho(\vec{r})\rho(\vec{r'})}{|\vec{r}-\vec{r'}|} - \int dr^3 V_{xc}(\vec{r})\rho(\vec{r}) + E_{xc}[\rho] \tag{2.17}$$

with the electron density $\rho(r)$ is calculated as Eq.2.18:

$$\rho(\vec{r}) = \sum_j |\psi_j(\vec{r})|^2 \tag{2.18}$$

The exchange-correlation term,

$$V_{xc} = \frac{\partial E_{\text{xc}}[\rho(\vec{r})]}{\partial \rho}, \tag{2.19}$$

has to be derived with approximations either analytically or numerically. There is formulation like, using the *uniform* electron gas model, as implemented in the Local-Density-Approximation (LDA) scheme. LDA can give accurate predictions for many cases [2]. Generalised Gradient Approximation (GGA) takes into account the density gradient and has better performance than LDA when the gradient of electron density becomes significant [19]. Accurate exchange-correlation functionals can be parameterised by means of Quantum Monte Carlo (QMC) sampling [20, 21] on the electron gas for a wide range of densities. In this thesis, the first-principles calculations were done with the GGA exchange-correlation functionals E_{xc}.

2.1.4 Blöch Theorem

In periodic system with periodic potentials: $V(\vec{r}) = V(\vec{r} + \vec{R}_n)$ where $\vec{R}_n$ are the lattice vectors, Blöch's theorem states that the electronic states $\psi(\vec{r})$ satisfies the following

condition:

$$\psi(\vec{r}+\vec{R}_n) = e^{i\vec{k}\cdot\vec{R}_n}\psi(\vec{r}) \tag{2.20}$$

The wavefunction therefore can be written as multiplication between a plane wavefunction $e^{i\vec{k}\cdot\vec{r}}$ and a periodic function $u(\vec{r})$ as follows,

$$\psi(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u(\vec{r}) \tag{2.21}$$

2.1.5 Brillouin Zone and K-points

Periodic boundary conditions (PBC) limit the eigenvector $\vec{k}$ in Blöch's scenario to discrete values:

$$\vec{k} = \sum_{i=1}^{3} \frac{m_i}{N_i} \cdot \vec{b}_i, \tag{2.22}$$

where $\vec{b}_i$ indicates the reciprocal vectors of the primitive lattice and $\{m_i\}$ take integer values. N_i is the number of unit cell along i-th basis directions. Therefore, $k \to 0$ as $\{N_i\} \to \infty$, which is the basis for the 'supercell' method. In a supercell calculation, the calculation performed at the Γ point (corresponds to $\vec{k} = 0$) upon an $N_1 \times N_2 \times N_3$ supercell is sufficient to yield all the electronic states for the system.

The electronic energy in the $\vec{k}$ space is discontinuous across the boundary : $|\vec{k}| = |\vec{k}+\vec{G}|$, where $\vec{G} = \sum_i m_i\vec{b}_i$. The different zones divided by the boundary are known as the Brillouin Zones. Monckhorst-Pack k-sampling is widely used in many *ab initio* code to implement the electronic states calculation with periodic boundary conditions. It generates sets of special points and sum of properties on these special points provides very good approximation to integration of the electron states over the entire or a portion of Brillouin zone [22].

The Blöch's scheme is for periodic crystals while for simulations like vacancies, defects or amorphous phase, a supercell with large enough range of interaction is usually needed to be constructed for the calculation. In order to study the crystal surface, we can apply a big enough vacuum to separate the interactions from its translational images.

2.1.6 Plane-Wave Expansion and Pseudo Potentials

When solving the KS equation (Eq.2.15) in periodic solid, plane-wave basis can be used to orthogonally expand the wavefunctions:

$$\psi_{n,\vec{k}}(\vec{r}) = \sum_{n} C_{n,\vec{k}} \exp\left[i(\vec{k} + \vec{G}_n) \cdot \vec{r}\right] \tag{2.23}$$

During a calculation, the plane-wave basis is usually truncated at a cutoff energy E_{cut}, corresponding to a sphere in the k-space centred at $\vec{k}$ vector, $\frac{1}{2}|\vec{k} + \vec{G}| \leq E_{\text{cut}}$.

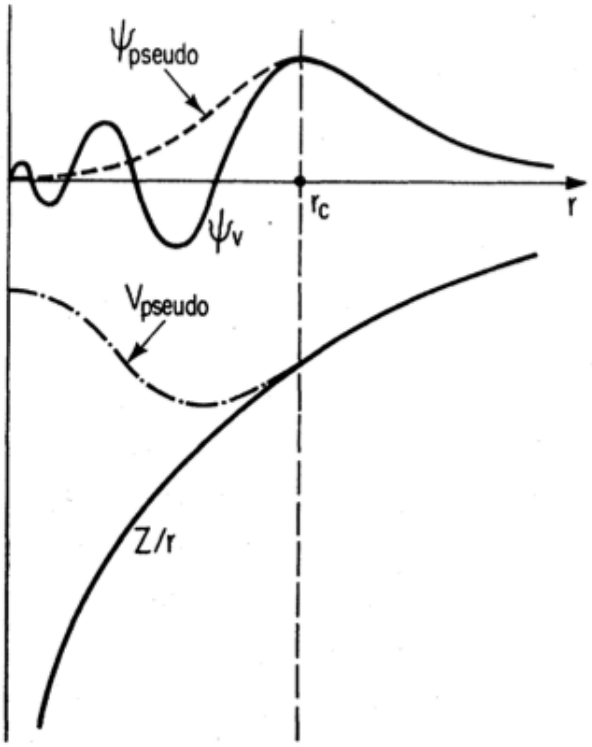


FIGURE 2.1: A plot illustrating the scheme of pseudo-potentials. Compared with the wavefunction (Ψ_{v}) from all-electron potential (Z/r), wavefunctions (Ψ_{pseudo}) from pseudo-potentials (V_{pseudo}) are smoothed close to the nuclei. r_{c} indicates the cutoff radius within which the pseudo potentials overtakes the real potential in pseudo-potential calculations. Reproduced from [23].

Pseudo-potentials are a methodology developed for the efficient performance of the plane wave expansion of the wavefunctions. As in the Fig.2.1, the real wavefunctions near the nuclei usually have strong oscillation compared to those in the outer space. The oscillation is largely caused by the kinetic energy gained near the nuclei and the requirement of orthorgonalisation with the core electron wavefunctions. For the oscillating area, large

numbers of plane waves have to be used for the convergence, which enormously increases the computation effort.

The pseudo-potential is introduced to some extent screens the attraction from the nuclei. The pseudo wavefunctions become smoothed within the cutoff radius but overlaps with the real potential for the outer space. Pseudopotentials are useful for investigating the valence electrons which are of principal interest for most of the cases and where the core electrons are typically not so important or can be recovered. There are commonly used ultra-soft pseudo-potential [24], norm-conserving pseudo-potential which are employed by *ab initio* packages like VASP [25, 26]. In this thesis, the ultra-soft pseudopotentials were adopted for the first-principles calculations using the VASP code.

2.1.7 Hellmann-Feynman Theorem

Due to the orthogonal properties of the eigen wavefuctions, $\langle\psi_i|\psi_j\rangle = \delta_{ij}$, the Hellmann-Feynman Theorem states that the QM force acting on ion $\vec{R}_I$ can be computed as,

$$\vec{F}_I = -\nabla_{\vec{R}_I} E = -\left\langle \Psi \left| \frac{\partial \hat{H}}{\partial \vec{R}_I} \right| \Psi \right\rangle \tag{2.24}$$

where E is the total energy calculated at ground state, and Ψ is the wavefunction for the N-electron system.

2.1.8 Phonon

Phonons are elementary excitation of the crystal lattice and are fundamentally related with a series of interesting properties or phenomena in material science, for instance, structural transformation [27], thermal conductivity [28], and super-conductivity [29]. At the first-principles level, phonons can be computed by supercell method [30]. Supercell methods based on the force constant matrix derived from Hellmann-Feynman force from finite atomic displacements (Eq.2.24) within a constructed supercell along high-symmetric directions.

Phonons at finite temperature T are distributed by the Bose-Einstein statistics (Eq.2.25) where multifold occupancy is allowed.

$$f(\epsilon_j) = \frac{g_j}{\exp(-\epsilon_j/k_BT) - 1} \tag{2.25}$$

In Eq.2.25, g_j corresponds to the degeneracy at energy level ϵ_j and k_B is the Boltzmann constant while T gives the temperature.

2.1.9 TB and DFTB

Tight binding (TB) was proposed by Slater [31] in the periodic crystal system and then extended to the atomistic configurations. The tight binding (TB) expands the electronic eigenstates semi-empirically with classical orbital basis.

$$\Psi_i = \sum_{v,\alpha} c_{vi}\phi_v(\vec{r} - \vec{R}_\alpha) \tag{2.26}$$

where $\vec{R}_\alpha$ indicates the centre position of α-th atom. There was later the extended non-orthogonal TB [32] using a non-orthogonal atomic orbital basis, which proved to have better transferability. In the TB schemes, we write the total energy as the sum of the band energy $E_{bs} = \sum_{i}^{occ.} < \Psi_i|\hat{H}|\Psi_i >$ and a repulsive energy E_{rep} part due to the repulsion from the electron pairs (Eq.2.27). The usual multi-body Hamiltonian $\hat{H}$ is thus replaced by a Hamiltonian matrix $H_{\mu v}$ and an overlap matrix $S_{\mu v}$ is formed by expansion with the atomistic basis $\phi_v(\vec{r})$ $(v = 1, 2, \cdots, N)$, as in Eq.2.28.

$$E_{\text{total}} = E_{bs} + E_{rep} \tag{2.27}$$

$$\sum_v (\hat{H}_{\mu v} - S_{\mu v}E_\mu)\phi_v(\vec{r}) = 0 \tag{2.28}$$

where overlapping matrix: $S_{\mu v} \equiv < \varphi_\mu|\varphi_v >$ and Hamiltonian matrix $H_{\mu v} \equiv < \varphi_\mu|\hat{H}|\varphi_v >$ and $\hat{H}$ indicates the multi-body electronic Hamiltonian operator. In the density-functional based tight binding scheme (DFTB), the atomistic basis $\{\phi_v(\vec{r})\}$ can be solved self-consistently in line with the KS equation with LDA/GGA exchange-correlation functionals. The calculations are performed upon the modified free atom model, where the extra repulsion term $(r/r_0)^N$ was found to be helpful for obtaining the diagonalised basis

set.

$$[\hat{T} + V_{\text{eff}}[\mathbf{n}_0(\vec{r})] + (r/r_0)^N]\phi_v(\vec{r}) = \epsilon_v\phi_v(\vec{r}) \tag{2.29}$$

where $\hat{T}$ indicates the kinetic operator, V_{eff} the effective potential as in the KS equation and $\mathbf{n}_0(\vec{r})$ is electronic density for the free atom model. The eigenfunctions $\{\phi_v\}$ are used as basis for constructing the TB wavefunctions Ψ as in Eq.2.28. The repulsive term E_{rep} in the total energy can be parameterised based on first-principles calculation results and the assumption of the density overlapping between electrons centred on different ions. Recent development on TB incorporated the self-consistent charge into the Hamiltonian matrix and has much improved performance for the ionic bonding system [33].

2.1.10 Summary

In this chapter, I have described the QM descriptions used in microscopic interpretation of the material properties. With the advent of supercomputing capacity, DFT has become a standard approach for the QM calculations of atomistic system up to hundreds of atoms, with the aid of the pseudopotential and plane-wave methodology. Further extending the scope of the investigation, however, meets the limitation arising from the $O(N^3)$ scaling. For large scale material systems, classical molecular dynamics usually has to be adopted instead, which will be described in the following part of this Chapter.

2.2 Molecular Dynamics

2.2.1 Introduction

Molecular dynamics (MD) has been widely used to explore the phase space of interacting particles at the harmonic regime or under conditions of external temperature or stress field [34]. One advantage of MD simulation is its capability to efficiently generate ensemble averages which can be linked with macroscopic observations. It is also a convenient tool to generate a dynamics trajectory in the configurational space.

Classical MD follows Newton's equation of motion,

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i \tag{2.30}$$

where m_i is the mass of the i-th particle, $\vec{r}_i$ the position vector. $\vec{F}_i$ is the force exerting on i-th particle and corresponds to the gradient of the potential energy surface,

$$\vec{F}_i = -\nabla_i V(\vec{r}_i) \tag{2.31}$$

where $V(\vec{r}_i)$ indicates the potential experienced by the i-th particle.

2.2.2 Ergodicity in MD

Ergodicity is a key issue in MD simulations and fundamentally determines if the correct ensemble averages can be obtained [35]. The *ergodic hypothesis* states that ensemble average $< A >$ over the phase space is equivalent to the time integral along MD trajectory, as in Eq.2.32.

$$< A >= \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{t}^{t+\Delta t} A(\tau) d\tau \tag{2.32}$$

In the above equation, $A(\tau)$ represents an atomic quantity at the MD time of τ.

2.2.3 Velocity-Verlet Algorithm

To evaluate the integration precision, a Taylor's expansion is performed to the position vector $\vec{r}$. This however, involves large errors and is only accurate to $O(\Delta t)$.

$$\begin{cases} \vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{\vec{F}(t)}{2m}\Delta t^2 + \frac{1}{3!}\frac{d^3\vec{r}}{dt^3}|_{t=t}\Delta t^3 + O(\Delta t^4) \\ \vec{r}(t-\Delta t) = \vec{r}(t) - \vec{v}(t)\Delta t + \frac{\vec{F}(t)}{2m}\Delta t^2 - \frac{1}{3!}\frac{d^3\vec{r}}{dt^3}|_{t=t}\Delta t^3 + O(\Delta t^4) \end{cases}$$

Verlet integration minimises the accumulation of error during the MD simulation by summing up the above two equations so that both the velocity term and third-order terms cancel out in numerical computations. The resulting $\vec{r}(t+\Delta t)$ is accurate to $\sim O(\Delta t^4)$ (Eq.2.33).

$$\vec{r}(t+\Delta t) = 2\vec{r}(t) + \vec{r}(t-\Delta t) + \frac{\vec{F}}{m}\Delta t^2 + O(\Delta t^4) \tag{2.33}$$

However, in the standard Verlet algorithm, the velocities are calculated by time average of the position and is accurate only to $O(\Delta t^2)$. The Velocity-Verlet Algorithm, also known as the 'leapfrog' algorithm [36] is used in the time integration for advancing the MD trajectory. As a development, this algorithm treats both velocity and position at the same precision which is accurate to the third-order in the Taylor's expansion. In Eq.2.34, I give expression for the velocity integration.

$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{1}{2}\left[\frac{\vec{F}(t)}{m} + \frac{\vec{F}(t+\Delta t)}{m}\right]\Delta t + O(\Delta t^4) \tag{2.34}$$

2.2.4 Thermostat

Under the Velocity-Verlet integration algorithm, the total energy is conserved within numerical precision, which yields the micro-cannonical (or NVE, which denotes constant particle Number, Volume, and Total energy) ensemble well by this means. Due to the practical significance of canonical ensemble (or NVT, which denotes constant particle Number, Volume, and Temperature), different approaches have been proposed to address the problem. However, to obtain a canonical ensemble, explicit approaches to simulate the constant temperature T have to be used.

- Early attempts to constrain the T usually included rescaling the velocity distribution artificially. The Anderson methodology was such an approach, where the velocity of a randomly chosen atoms are rescaled to that from the Maxwell distribution at the target temperature [37]. This rescaling scheme however, usually causes dramatic changes to the MD trajectory. As an improvement, the Berendsen thermostat adopts smooth rescaling of the instantaneous kinetic energy towards the target kinetic energy [38].

- Another approach is by adding stochastic contribution into the MD process to stimulate the ergodicity, such as Langevin dynamics [39], as expressed in Eq.2.35 where $\vec{P}$ indicates the momentum and V the potential energy,

$$\dot{\vec{P}} = -\nabla V - \gamma \vec{P} + \vec{R}(t) \tag{2.35}$$

  It is carried out stochastically by considering a damping force $\gamma\vec{P}$ and a random force $\vec{R}(t)$ which has only short-term correlation along time,

$$\left\langle \vec{R}(t), \vec{R}(t + \Delta t) \right\rangle = \delta(\Delta t) \tag{2.36}$$

  Based on the Stokes-Einstein fluctuation-dissipation relation, $\vec{R}$ has a Gaussian distribution for time increment of Δt while the variance can be derived as:

  $\sqrt{2m\gamma K_B T/\Delta t}$.

  In the Langevin scheme, the strength of the thermostat can be adjusted by means of the parameter γ. By use of optimally chosen γ, fast convergence for the ensemble average can be achieved.

- A widely used approach is by introducing a thermostat to couple the systems with an external heat bath. This way can be applied to generate reliable MD trajectories. The Nóse thermostat by rescaling the time of the subsystem to obtain the correct NVT partition function and thus ensure the NVT conditions[40]. However the deterministic nature of the thermostat can not assure ergodic condition for the ensemble either and therefore cannot produce reliable ensemble averages. Combining of the Nóse and Langevin thermostat in practice was also shown to have good performance [41].

2.2.5 Classical Force Fields

Classical potentials describe interactions in the system by function parameterisation using either experimental or *ab initio* calculation results. Due to their computational efficiency, atomistic systems containing millions of particles are able to be investigated. Classical potentials have been widely used over the years and found applications for studying bulk defects, vacancies or amorphous phases, etc. More than that, they also have extensive applications in simulating bio-chemical materials.

In material simulations, there are a number of commonly used classical potentials, such as Lennard-Jones potential, Stillinger-Weber potential, bond-order Tersoff potentials and environment-dependent inter-atom potentials (EDIP), *etc.* In the following, taking Silicon as an example, I will give a review of them.

2.2.6 Lennard-Jones Potentials

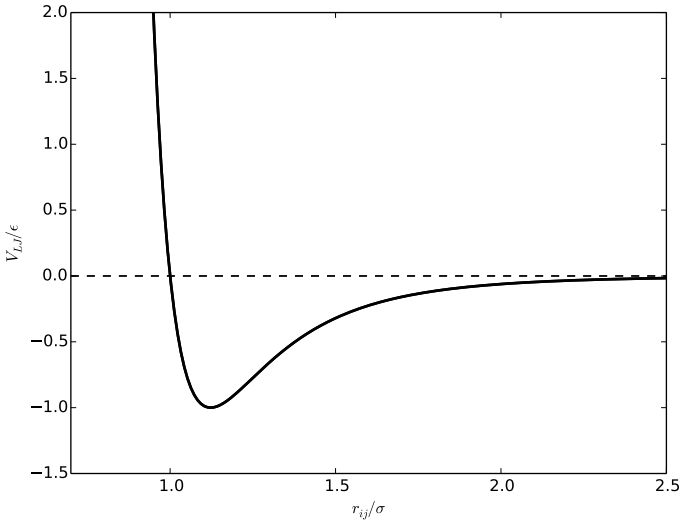


FIGURE 2.2: A schematic plot illustrating the LJ potential (black solid line) with respect to the pair distance r_{ij}.

The LJ potentials are comprised of repulsion and attraction terms. One popularly used form is given in Eq.2.37, where $(\sigma/r_{ij})^{12}$ is adopted for the repulsion while $(\sigma/r_{ij})^{6}$ for the attraction, the latter typically in agreement with decaying of the van der Waals interaction [42],

$$V_{LJ} = 4\epsilon \cdot \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2.37}$$

In the equation, ϵ corresponds to the depth of the potential well and σ indicates the pair-distance where the potential energy equals zero on the left repulsion region in Fig.2.2.

LJ potentials provide insight into many physical properties, especially useful in the case that there is no theoretical framework to be referenced to [42]. They found application in modelling of liquid, gas or metals, but due to the simplicity of functional form and absence of three-body interaction, their predictive capability is very limited. For example, they can not predict the stable phase of diamond structure in semiconductors such as silicon.

2.2.7 Stillinger-Weber (SW) Potentials

The Stillinger-Weber potential was one of the first attempts to describe covalent bonding in semiconductors by incorporating both two-body V_2 and three-body angular interaction V_3 terms, as expressed by the following equation:

$$V_{\text{sw}} = \sum_{i<j} V_2(r_{ij}) + \sum_{i<j<k} V_3(r_{ij}, r_{ik}, \theta_{ijk}) \tag{2.38}$$

where θ_{ijk} indicates the angle between the bonds of r_{ij} and r_{ik}. The two terms can be explicitly written as:

$$V_2(r_{ij}) = \epsilon \cdot f_2(r_{ij}/\sigma) \tag{2.39}$$

$$V_3(r_{ij}, r_{ik}, \theta_{ijk}) = \epsilon \cdot g\left(\frac{r_{ij}}{\sigma}\right) g\left(\frac{r_{ik}}{\sigma}\right) h(\theta_{ijk}) \tag{2.40}$$

where,

$$f_2(r) = (Ar^{-4} - B) f_{\text{cut}}(r) \tag{2.41}$$

and the angular function has following forms,

$$h(\theta_{ijk}) = \lambda(\cos\theta - \cos\theta_0)^2 \tag{2.42}$$

The parameters A, B and λ are obtained by fitting the potential to material properties measured by experiment or calculated using first-principles methods. The cutoff function in Eq.2.41: $f_{cut}(r)$ is set 0 when $r > \sigma r_c$. To describe the sp^3 bonding type in diamond structure, θ_0 is set to be 1/3. The radial function $g(r_{ij})$ also incorporates a cutoff function which monotonously decreases as the bond length r_{ij} stretches.

For Silicon, by fitting the parameters to reproduce the diamond structure as the most stable phase, the SW potential yielded satisfactory thermodynamic properties for both bulk and liquid Si and it also predicts the approximately correct melting temperature [43]. By introducing parameters for the two- and three-body interaction terms between Silicon (Si) and Hydrogen (H) , SW potentials were recently extended to modelling of the interactions between H and Silicon surfaces [44]. Because SW potentials have an analytical form close to that required by the Harrison condition [45], good agreement with experimental and DFT results were also found for the elastic properties of diamond structure [46]. In this thesis, SW potentials are used to calculate the classical force vectors that are used to augment the representation of the atomic environments, which will be presented in Chapters 5 and 6.

2.2.8 Tersoff and Brenner Potentials

The Tersoff potential was one of the reliably used empirical potentials to study material properties like lattice dynamics, point effects or amorphous phases [47, 48]. The idea is by deriving a potential based on the concept of bond order. It expands the potential function with local bond order parameters $\{r_{ij}\}$ as in Eq.2.43.

$$V_{ts} = \sum_{i<j} f_c(r_{ij})[AV_2^{(A)}(r_{ij}) - B_{ij}V_3^{(R)}(r_{ij})] \tag{2.43}$$

In the equation, $f_c(r_{ij})$ corresponds to a cutoff function for the interacting atoms, and $A, \lambda_1,$ and, λ_2 are parameters to be fit. The first term $V_2^{(A)}$ takes similar form as the two-body interaction in SW potentials while the second term $V_3^{(R)}$ indicates the repulsion from three-body interaction, which however, within the Tersoff potentials is a function of the local bond order.

The parameter B_{ij} indicates the weight of the repulsion contribution competing with the bonding between atom pairs i and j. It depends on the local bond order environment and at first-order approximation can be expressed as function of the local coordination number (Z): $b_{ij} \propto Z^{-1/2}$.

In 1990, based on the Tersoff potential, Brenner potentials were introduced to incorporate radial contributions to the bond order B_{ij}. Brenner potentials have much improved performance for hydrocarbon systems [49].

2.3 First-Principles Molecular Dynamics

Applying a highly accurate electronic description into molecular dynamics is implemented by the First-Principles Molecular Dynamics (FPMD), which is widely used to explore the quantum chemistry processes. Such a description is realised in QM chemistry with the capability to explore the configuration space leading to the unlimited transferability [50].

Among the FPMD schemes, the Car-Parrinello molecular dynamics (CPMD) was the first approach to consider the dynamics within framework of the Lagrangian equation including electronic degrees of freedoms [51], as expressed as follows:

$$L = \sum_i \frac{1}{2}\mu \int_\Omega d^3r|\dot{\psi}_i|^2 + \sum_I \frac{1}{2}M_I\dot{R}_I^2 + \sum_\nu \frac{1}{2}\mu_\nu \dot{\alpha_\nu}^2 - E[\{\psi_i\},\{R_I\},\{\alpha_\nu\}] \tag{2.44}$$

where L indicates the Lagrangian of the system, $\{\vec{R}_I\}$ the ionic coordinates, M_I the mass of the I-th ion, ψ_i wavefunction occupied by the i-th electron, and the first-term and second term correspond to the kinetic energy of the electrons and ions respectively. The last term $E[\{\psi_i\},\{R_I\},\{\alpha_\nu\}]$ indicates the potential energy, which comes from solely the Schrödinger equation as discussed in section 2.1.

The equations of motion can be derived as :

- For the ion:

$$M_I\ddot{R}_I = -\nabla_I E \tag{2.45}$$

- For the electronic degree of freedom :

$$\mu\ddot{\psi}_i = -\delta E/\delta\psi_i^* + \sum_k \Lambda_{ik}\psi_k \tag{2.46}$$

- For the external degree of freedom:

$$\mu_\nu\ddot{\alpha}_\nu = -\partial E/\partial\alpha_\nu \tag{2.47}$$

Another FPMD scheme is the Born-Oppenheimer molecular dynamics, where the dynamics of the ionic coordinates is carried out by Eq.2.45 with $-\nabla_I E$ referring to the Hellman-Feynnman forces. The electronic degrees of freedom are relaxed to its ground

state before we advance the MD trajectory of the ionic coordinates. Compared with the CPMD, electronic motion equation in Eq.2.46 will be replaced by solving the standard Schrödinger equation given the ionic coordinates $\{\vec{R}_I\}$. BO MD is an efficient way to implement the DFT static calculations into the molecular dynamics and is used as good approximation to the accuracy at first-principles level as long as the BO approximation holds. In our machine learning of quantum forces (Chapters 4-7), I will show that useful Bayesian inference from the database containing Hellman-Feynnman forces can significantly accelerate the computation efficiency of BO molecular dynamics.

2.4 Beyond the classical calculations

Molecular dynamics at the classical and first-principles levels have been introduced and both of these approaches are used widely targeting different domains of problems. With DFT accuracy, the behaviour of materials can be understood and predicted reliably from the interaction due to the electrons, while the classical approaches extends the simulation scale enormously and practically narrows the gap between theoretical description and experimental observations. However, both approaches have limitations when it comes to multi-scale simulations. The classical potentials are less suited to use in new chemical environments beyond the fitting properties and the first-principles methods have limited applicability typically due to its $O(N^3)$ scaling factor, arising from the need to keep wavefunctions orthogonal in the iterative numerical procedures. Linear scaling DFT was recently developed and can accelerate the DFT calculations, but also has limited applications [52, 53].

Beyond the classical description, methodologies such as QM/MM [11], 'learn-on-the-fly' (LOTF) MD [12, 54], have been developed and successfully employed into the multi-scale simulations, for instance, to model the fracture, defects or dislocations, to be reviewed in the following sections.

2.4.1 QM/MM Embedding

QM/MM embedding was first proposed by Warshel and Levitt in 1976 to simulate the reaction taking place between the enzymes and substrates together with the surrounding solvent [55]. The long-range electrostatic interaction between the enzyme and subtract

as well as the polarisation energy of the environmental water has to be treated with classical potentials. The bond cleavage and charge redistribution in the subtract however were evaluated with QM calculation. The embedding of QM/MM for the entire system developed in this work was later widely used in simulations both for bio-chemical system [56] or multi-scale physics system [11]. Due to their excellent work in this area, along with Karplus, Warshel and Levitt were awarded the Nobel prize for Chemistry in 2013. In literature [56], the application of the QM/MM embedding into the enzymatic reactions (such as in methane monooxygenase) were reviewed and among these systems it is explained that the coupling between two distinct potential energy surfaces using QM and MM were challenging to address.

For simulation of multi-scale system like fracture, atomic bonding behaviour or chemo-mechanical process near the crack tip exhibit much scientific interest under the stress from the loading, while the vast majority surrounding atoms stays within roughly the same area of configurational space throughout one simulation. QM treatment is thus assigned to the region where the electronic behaviour is of significant interest, while for the less-pertinent environmental configurations, molecular mechanics (MM) is employed during the calculations. The two parts are coupled in the way that the MM part provides a long range stress field which effects the inner QM atoms substantially while the accurate configuration of the QM part, in return, determines the dynamics of the outer part. In a QM/MM calculation for Silicon fracture, around 100 atoms in the inner part are treated with QM and embedded with the outer MM part via a buffer region. The QM part is thus greatly reduced to the range accessible to the full DFT calculation. Different QM/MM hybrid schemes have been proposed to treat the chemically active part to QM accuracy, while the less relevant part are evaluated with the MM description [11, 46].

- The **energy mixing** scheme. One popular approach is called ONIOM [57]. This mixing scheme is performed based on a total energy written as,

$$E_{tot} = E_{sys}^{MM} + E_{cluster}^{QM} - E_{cluster}^{MM}$$

  In the above equation, E_{sys}^{MM} the total energy of the entire system calculated with MM approach, $E_{cluster}^{QM}$ the passivated cluster energy calculated with QM, and $E_{cluster}^{MM}$ indicates the energy for the passivated cluster which are calculated with

MM method. The dangling bonds for the carved cluster are usually passivated by hydrogen or pseudo-atoms upon the boundary atoms. Forces are accordingly calculated as the derivative of the energy formula in Eq.2.48 during the simulations.

- The **force mixing** scheme. A boundary region is usually marked to connect the two descriptions, i.e. QM and MM, using a smooth transition parameter : λ, the force in the boundary regions takes the form for instance: $\vec{F}_\lambda = \lambda\vec{F}_{\mathrm{QM}}+(1-\lambda)\vec{F}_{\mathrm{MM}}$, where the parameter λ moves from 0 to 1 for forces on atoms from the MM zone till the QM zone. LOTF MD (to be discussed in next section) also adopts force embedding and QM forces are calculated for carved clusters with buffer region, but differs from other force mixing in that the inaccurate QM forces in buffer part are discarded when mixing with the MM region. The dynamics of the system is however, advanced by an adjustable potential upon the MM forces, incorporating the information of the QM forces, while the momentum conservation is exerted for the entire system in the mixing [11].

- The **electrostatic mixing** scheme. For the long-range Coulomb interactions such as in the systems of Silica, Silicon-Carbide and ever more importantly for biological systems, the mixing between two descriptions are more challenging. In this case, a proper embedding should prevent the electron density escaping from the QM region to the MM region, which is the so-called 'spilling out' effect. Laio et al. proposed a scheme to address that issue by introducing a Hamiltonian term explicitly coupling the Coulomb multi-pole interaction between the QM charge distribution and the MM points charges thus the interactions between the QM atoms and the distant MM atoms can be modelled in MD simulations [58].

2.4.2 LOTF Molecular dynamics

The 'Learn-on-the-fly' (LOTF) method proposed by De Vita and Car in 1998 [54] addresses the mismatching between QM and MM descriptions and the transferability problem met when using empirical interatomic potentials in molecular dynamics.

In the LOTF MD, a classical force field is augmented by a simple, adjustable potential $V_{\mathrm{adj.}}$, whose parameters can be updated with newly computed QM forces during the simulation [12]. The overall potential that is informed by QM calculation results and

used for advancing the MD trajectory is:

$$\mathrm{V_{LOTF}}(\mathrm{R}, \alpha) = V_{\mathrm{MM}}(\mathrm{R}) + V_{\mathrm{adj.}}(\mathrm{R}, \alpha) \tag{2.48}$$

where V_{MM} indicates the classical potential, for instance, the Stillinger-Weber Potential [43] when simulating Silicon. $V_{\mathrm{adj.}}(\mathrm{R}, \alpha)$ corresponds to the parameterised adjustable potential.

$$V_{\mathrm{adj.}}(\mathrm{R}, \alpha) = \sum_{i<j} \alpha_{ij} \cdot R_{ij} \tag{2.49}$$

One of the most recent implemented parameterisation form for $V_{\mathrm{adj.}}$ takes the form of Eq.2.49. In the equation, R_{ij} indicates the bonding distance between atom-i and atom-j while $\{\alpha_{ij}\}$ are the parameters to be fit to the QM results. Based on the potentials in Eq.2.48, forces are derived as the negative gradient of the potential energy, and are written as:

$$\vec{F}_{\mathrm{LOTF}} = \vec{F}_{\mathrm{MM}} + \vec{F}_{\mathrm{adj.}} \tag{2.50}$$

Regarding the QM fitting part, the adjustable parameters of the potential are dynamically optimised by minimising the discrepancy between QM forces and MM forces at an interval of n-step MD simulation. The minimisation is expressed by the following equation:

$$\min_{\{\alpha\}} \|(\vec{F}_{\mathrm{MM}} + \vec{F}_{\mathrm{adj.}}) - \vec{F}_{QM}\| \tag{2.51}$$

The optimised parameters $\{\alpha\}$ are used to make calculations for the next cycle of n-step MD simulation. At the end of this cycle, a new QM calculation is performed and the parameters set $\{\alpha_{ij}\}$ is again optimised and updated. The MD is carried out in such a way that the computationally expensive QM calculations are only performed at every n simulation steps and therefore, the overall speed of the calculation is accelerated straightforward by a factor of n compared to the full QM calculations.

Predictor-corrector algorithms are employed in the standard LOTF MD simulation to make best use of the optimised potential V_{LOTF}, which I will explain in the following. We know that the predictor cycle is an n-step run with the updated parameter set $\{\alpha_1\}$. At the end of each predictor cycle, QM force are calculated and the parameter set are refitted to be $\{\alpha_2\}$. A corrector is a recalculation from the initial configuration of the

predictor cycle but with V_{LOTF} which is interpolated using two parameter set obtained at two successive QM calculation points, i.e. $\{\alpha_1\}$ and $\{\alpha_2\}$. The interpolated parameter set is :

$$\alpha = \lambda\alpha_1 + (1-\lambda)\alpha_2 \tag{2.52}$$

where index λ runs from 0 to n across the corrector cycle.

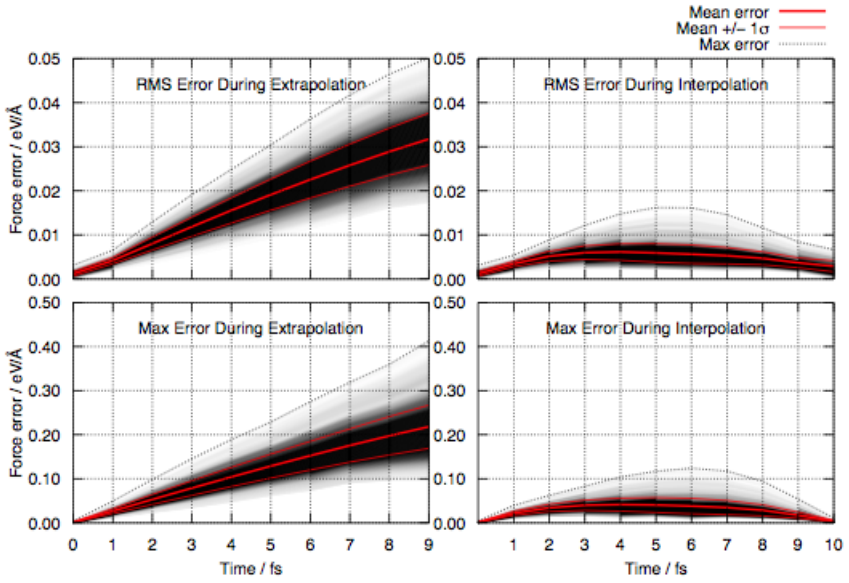


FIGURE 2.3: Reproduced from [46]. LOTF Force errors during the predictor and corrector cycles from a number of independent MD runs, the RMS error were marked by the red lines. The test system is 64-atom bulk Si at 2000 K with the QM Hamiltonian under the DFTB framework.

A systematic plot about the predictor and corrector errors during the LOTF MD is plotted in Fig. 2.3. During the predictor cycles, un-updated potentials were used to do the force calculation therefore, the error shows linear dispersion from the benchmark of DFT forces. The predictor error reaches a maximum at the end of the cycle, where in the LOTF scheme, new QM forces are computed and the potentials are updated with the new set of parameters $\{\alpha\}$.

The corrector cycle is a recalculation of the predictor cycle from the same starting configuration but with now the updated V_{LOTF} potential. The error therefore has its minimum at the beginning and end points of the cycles where the QM fitting was performed, while the maximum at the middle point of the cycle. The end configuration during the predictor cycle however are slightly different from the predictor cycle so care should be taken not to extrapolate too far outside the domain. In Fig.2.3, the error

curves corresponding to the test on the trajectories at independent MD runs, and the 10 step predictor-corrector calculation in most of the tests assured a chemically desirable accuracy that is below 0.1 eV/Å, and a factor of $3-4$ times smaller than that in the predictor cycles.

2.4.3 Summary

LOTF MD drops the energy conservation of other QM/MM approaches and instead, it enforces the force toward the QM accuracy by use of adjustable potentials. Using LOTF MD together with the embedded QM/MM, Kermode et al. successfully investigated the low-speed fracture of Silicon and correctly described the brittle nature of the fracture system [13]. A key advantage of LOTF MD is that the QM region can move in the on-the-fly way, unlike the conventional QM/MM embedding.

The limitation of these methods is however, that the accuracy can only be guaranteed on the condition that a good classical potential is available beforehand while as we know, the derivation of good classical potential is usually a demanding task both in terms of physical intuition and skills. Also, even during the interpolation corrector, accuracy diverges very fast as the complexity of the configurations becomes broad on range and the method is usually valid for a limited amount of simulation steps ~ 10. These methods are useful where the complexity is localised.

In Chapters 4-7, I will introduce a new approach aiming for the use in large-scale molecular dynamics, where force are predicted by Machine-learning (explicitly, Gaussian Processes) from QM database which is updated in an 'on-the-fly' fashion. Typical advantage of this methodology is that, no parameterisation is involved in this learning scheme, thus the prediction and learning process becomes valid in a broad range of structural variation. In this methodology, minimal amount of QM calculations are called for and large steps of interpolation and extrapolation can be achieved in the dynamics, and the boundary problems associated with the QM/MM embedding are naturally lifted. Before introducing the ML of force calculation in next Chapter, I will present a model to study the typical multi-scale system, e.g. fracture, which is one of the most interesting playgrounds for all these developed algorithms.

Chapter 3

Fracture Modelling

3.1 Introduction

In this section, I will introduce some general background knowledge relevant to the fracture simulations. Fracture is the lifetime limiting failure mode of many materials and is of tremendous technological concern, from mining to ceramics and the glass. The fracture in materials can be divided into two broad classes: ductile and brittle fractures. For the former, enormous plastic deformation accompanies the fracture process while for the latter, the crack propagates along energetically-favoured cleavage planes [59, 60].

Brittle fracture as one of the most typical multi-scale problems has attracted much research interest, especially recently, from physicists in the computational material science community [13]. In this kind of fracture system, bonding events in a concentrated area and stress field from long-range distance comes into a mutual play. The long-range interaction nature requires the incorporation of a number of atoms which can only be dealt with a classical approach, while for the short range chemical related events, DFT is most desirable for an accurate description.

3.2 Griffith's Criterion

The first theoretical work on fracture was carried out by Griffith in 1921 from the point of view of thermodynamics [61]. Crack propagation involves two processes: creation of new crack surfaces and release of elastic energy due to the applied stress field. Suppose

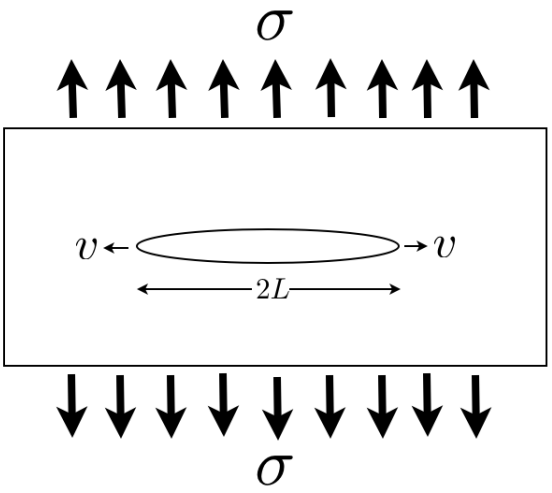


FIGURE 3.1: A schematic plot showing the theoretical model in fracture. σ is the applied stress field, v the propagation velocity and $2L$ the crack length.

the crack propagation by a length of dL, the released elastic energy is dE_c and the energy consumed to create new surfaces dE_s. The following inequality must be satisfied for the crack to propagate,

$$\frac{dE_s}{dL} + \frac{dE_c}{dL} \leq 0 \tag{3.1}$$

For crack model with the specific geometry illustrated in Fig.3.1, the above energies are calculated as:

$$\begin{cases} E_s = 4\gamma L \\ E_c = -\pi L^2 \sigma^2 / E' \end{cases}$$

where L is the crack length, γ the surface energy density and E' is the effective Young's Modulus defined as follows:

$$E' = \begin{cases} E & \text{for plane stress} \\ E/(1-\nu^2) & \text{for plane strain} \end{cases}$$

The effective Young's modulus is the usual Young's Modulus E for in-plane stress and $E/(1-\nu^2)$ for the plane strain case with ν the Poisson ratio.

Substituting the energy expression into the Griffith's relation in Eq.3.1, we obtain the criterion for the crack to propagate. The applied stress σ should exceed the critical loading stress σ_c expressed as follows:

$$\sigma_c = \sqrt{2E'\gamma/\pi L} \tag{3.2}$$

In the study of crack propagation, an elastic energy release rate to the crack tip is usually introduced,

$$G = -\partial E_c / \partial L \tag{3.3}$$

Using critical elastic energy release rate G_c, the Griffith criterion can be expressed as,

$$\mathrm{G}_c = 2\gamma \tag{3.4}$$

which means that, the critical loading is the twice the *surface energy density*. This is a more generally used form in the community of fracture research. However, the critical energy release rate G_c for the crack to propagate in both experiment and atomistic simulations are typically higher than the prediction by the Griffith criterion at the continuum limit (Eq.3.4). This is attributed to the energy barrier for bonds to break at the atomistic level, which is known as *lattice trapping*. The concept was introduced by Thomson and Rana from their analytical model [62].

3.3 Velocity of the Crack Propagation

According to linear elastic fracture mechanics (LEFM) for semi-infinite crack model [63] where a linear relation is adopted between the stress and strain, the velocity of the crack under loading G is given by :

$$v = c_R \left[1 - \frac{\Gamma(v)}{G}\right] \tag{3.5}$$

In the above equation, c_R is the Rayleigh wave speed, equal to the velocity of acoustic surface waves, $\Gamma(v)$ the velocity-dependent fracture energy which is approximately equal to the Griffifth's critical loading G_c at low speed crack regime, and G is the strain energy release rate defined as before [13].

There is a discrepancy between this theoretical prediction and experimental measurements. For instance, typically from Eq.3.5, the maximum velocity is the Rayleigh wave speed c_R while the maximum velocity observed in experiment is usually less than the Rayleigh wave speed, about $\sim (20\% - 80\%)c_R$ [64]. The explanation for this discrepancy lies in the crack instability above some critical velocity [65]. Models that additionally consider the phonon dissipation energy in the crack propagation improved their agreement [66].

Classical Approach. Different classical approaches were among the early attempts to address the fracture simulations at the atomistic scale. With potentials such as SW, TS, and EDIP, Silicon fracture however is incorrectly predicted to be ductile [67–69]. This discrepancy between these predictions and other accurate atomistic simulation as well as experiments are largely due to the fact that stress concentration diverge as $\sim 1/\sqrt{r}$ near the crack tip, leading to anharmonic bond activity that is barely captured by the classical potentials [59, 69, 70].

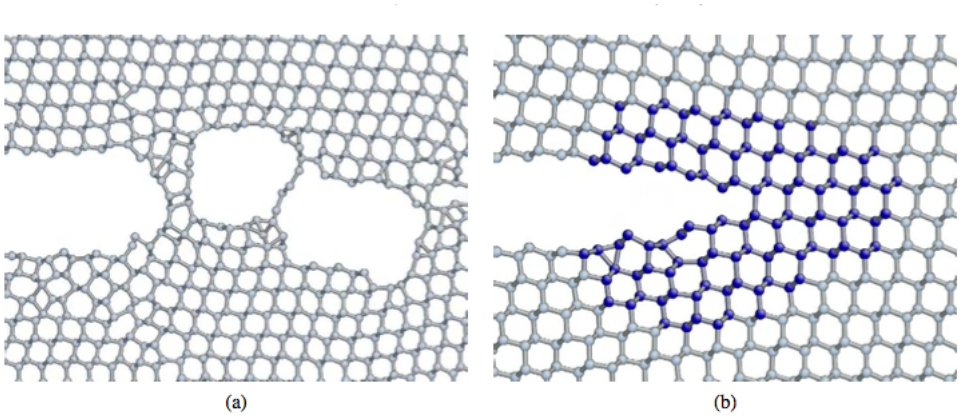


FIGURE 3.2: In panel (a), the Silicon fracture profile from a classical trajectory calculated uniformly using the SW potentials. The crack surface is typically observed to be ductile. In panel (b) gives the snapshot of the crack from the QM/MM embedding scheme. The Figure was reproduced from [59].

Regarding the crack propagation velocity, there was a long-standing yet unconfirmed prediction that under continuously increasing loading, the crack start propagating only at a finite velocity v_0. The forbidden velocity band between 0 and v_0 is called the 'velocity gap', which was ever reported in the experiment work of [71, 72]. However, there was also other experimental work suggesting that no sign of the 'velocity gap' were actually observed [73]. These distinct results make the velocity gap a topic of debate in the fracture community. A systematic study of this issue was thus motivated and described below.

Simulations of Silicon fracture were performed by my collaborators with a machine-learning potential: GAP potentials [1] [74]. Before application, this GAP potential was carefully trained with 400 reference configurations that were sparsified from a database of ∼1500 bulk configurations and ∼1500 configurations from the (111) fracture surface. The crack in these simulations, was performed on the (111) cleavage plane of Si crystal,

[1] A detailed description of the ML potentials and GAP is found in Section.4.2 of this thesis.

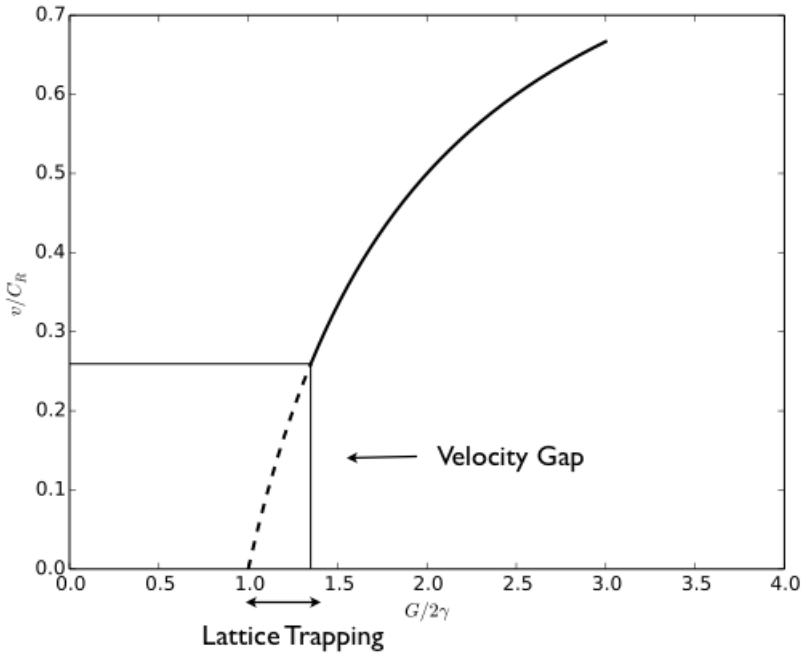


FIGURE 3.3: A schematic plot illustrating the velocity gap. The dotted line and the rectangle indicates the forbidden velocity region: $[0, v_0]$ in the fracture velocity curve with respect to the loading rate. The range of the velocity gap in the plot is arbitrary and can varied depending on the specific material.

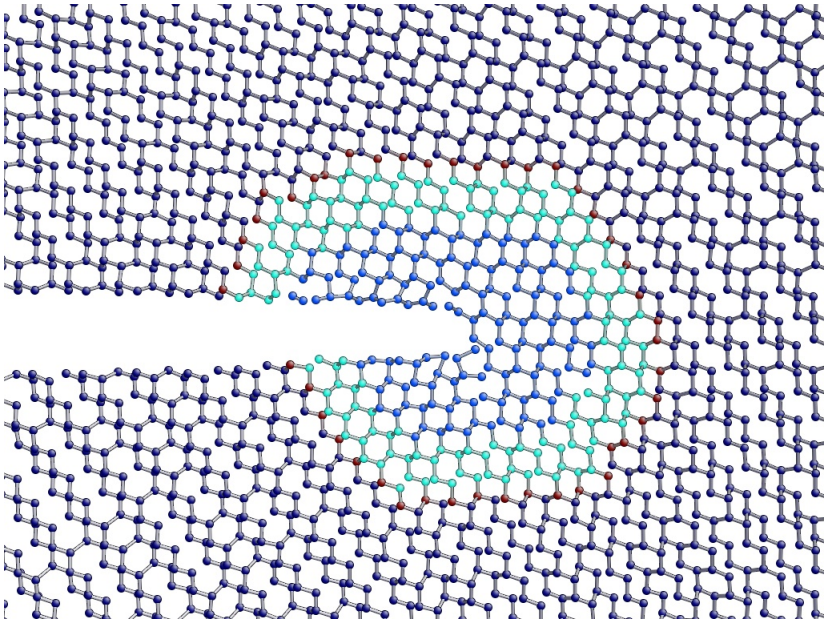


FIGURE 3.4: A snapshot near the Si crack tip at temperature of 300 K with different colours marking the GAP part (inner region, light blue), the MM part (outer region, dark blue) by SW potential and the buffer region between them (intermediate region, cyan), respectively. This snapshot was from the crack simulation within a LOTF GAP/SW model, where GAP calculation replaces the QM calculations in the conventional QM/MM embedding.

while the crack propagates in the $[11\bar{2}]$ direction, under several different temperatures from 5 K to 500 K.

Though hugely advantageous in terms of efficiency compared to the DFT, dealing with a system with 111,000 atoms on the fracture system, the cost to use full GAP potential is still prohibitive, 100 $\sim$ 1000 times more expensive than the SW classical potential, with the dominant cost on the calculation of the descriptor for the atomic configurations (details in section 4.2.3). GAP/MM embedding was used in the simulations. The chemically active region near the crack tip was calculated using GAP potential and the long-range stress concentration is calculated with the SW potentials. A snapshot of the embedded GAP/MM scheme was shown in Fig.3.4, with a buffer region to obtain the correct GAP forces on the inner part and discarded during the mixing with MM forces (see Section. 2.4.1).

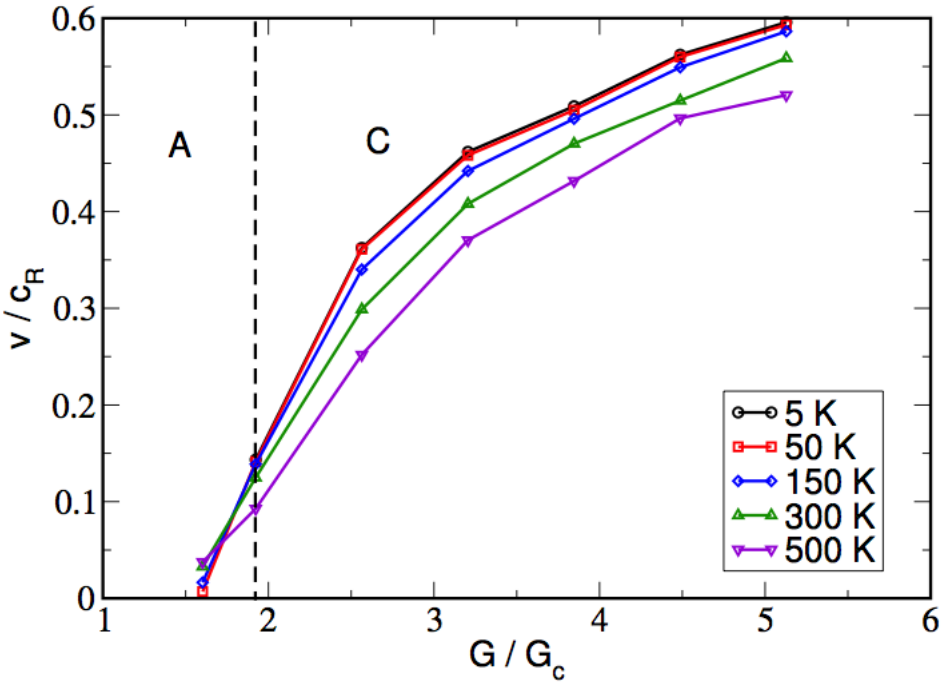


FIGURE 3.5: The crossover plot of crack under loading at different temperatures. The velocities for the lowest energy release rate $G = 2.5$ J/m^2 were from process of slowing down a running crack. In the Figure, **A** and **C** refer to thermally activated and catastrophic region, respectively. The plot is from [74].

As one way of validation, the use of GAP potential was verified on the brittle surface of the crack and reproduced the Pandey surface reconstruction of 5- and 7- atoms rings, which is formed due to the energetically-favoured π-bonding type on the (111) surface

[75]. This reconstruction was also reported using QM/MM embedding calculation with the QM forces calculated within the DFT framework. [13].

The velocities under different energy release rates G and temperatures T were explored, as shown in Fig.3.5. For the low loading part, thermally activated fracture modes were revealed (denoted as **A** in Fig.3.5) and is attributed to the lattice trapping. The higher kinetic energy favours the overcoming of the energy barrier due to the lattice trapping and thus higher propagation velocity. Above a critical loading rate, higher kinetic energy starts slowing down the crack propagation, and this renders a crossover in the plot in Fig.3.5. This effect becomes much pronounced in the catastrophic regime (**C** in Fig.3.5). An autocorrelation calculation of the bond breaking between each bond-breaking sites along the propagation direction was performed for the different temperature cases and only very weak autocorrelation was ever found, suggesting the relatively independent breaking process for each sites [74]. In the following section, I describe a simple model approximate the bond breaking process near the crack tip and further understanding of the mechanism is possible.

3.4 Results: Brittle Bond Breaking

During the atomistic simulation using the GAP/SW embedding method, the reversed contribution from temperature to the crack propagation velocity was found for the thermally activated and catastrophic loading regions. The GAP potential in the simulation has no chemical accuracy confirmed (see section 4.2.3) and thus the question is still open as to the bond breaking process. In the case of no experimental observation available to confirm about the findings, a theoretical model is thus designed to provide a different perspective to the underlying mechanism, which motivated the work to be described below.

In this simplified atomic model with LJ interaction potential as depicted in Fig.3.6, we ignore the site correlation of bond breaking near the crack tip in the crack propagation directions, which is appropriate according to the result in GAP/SW. Pulling forces exerted on the edge atoms are used to atomically simulate the stress fields near the crack tip and separation of the bonding structure is carried out at a relative velocity of $\mathbf{v}$. Due to the general atomistic features, this model is not restricted to Silicon and is

designated to address the description for bond breaking in a group of brittle fracture system. As one of the significant factors affecting the bond breaking, the stiffness of the back bonds compared to the central bonds was also explored with the stiffness ratio defined as the proportion of the bond energy between the back bonds and the central bond. We adopted uniform bonding which corresponds to the stiffness ratio of 1.0 as shown in panel (a) of Fig.3.6 and triple back bonding with stiffness of $\sim 4/3$, as in panel (b), in analogy to the tetrahedral bonding type in Silicon.

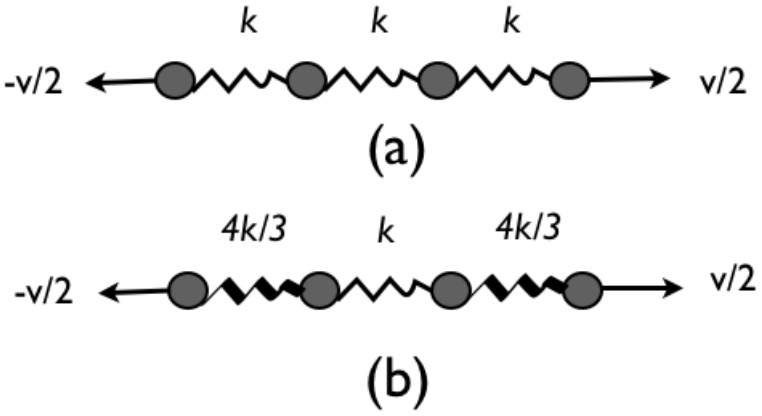


FIGURE 3.6: A schematic plot showing the bond-breaking model with different back bond strength ratio with respect to the central bonding as illustrated in (a) and (b), respectively. The edge atoms in both case are considered to be moving with constant velocities.

At $T = 0$, classical dynamics of the LJ interaction system was performed for each initialised separation velocities $\mathbf{v}$ until bond breaking which is defined to happen whenever the pair atomic distance exceeds 3 times of the equilibrium distance. The external work W consumed to break (any) one of the bonds was calculated against each of the separation velocity $\mathbf{v}$ under the energy conservation law by integration of force contribution along the separation process. This work W averaged over sampling of the possible variables (e.g. phonon vibration) is thus connected with the propagation velocity of crack tip through each of the perpendicular bonds in real crack system. Results for the consumed work W against the separation velocity $\mathbf{v}$ is plotted in Fig.3.8. The velocity is given in units of the sound speed $\mathbf{v}_s$ for this LJ dimer (pair atoms with LJ interaction) calculated as,

$$\mathbf{v}_s = \sigma/T_{LJ} = 6\sqrt{2\epsilon/m} \tag{3.6}$$

where σ and ϵ are the typical LJ potential parameters and m indicates the reduced mass of the LJ dimer, T_{LJ} the phonon vibrating period for the dimer.

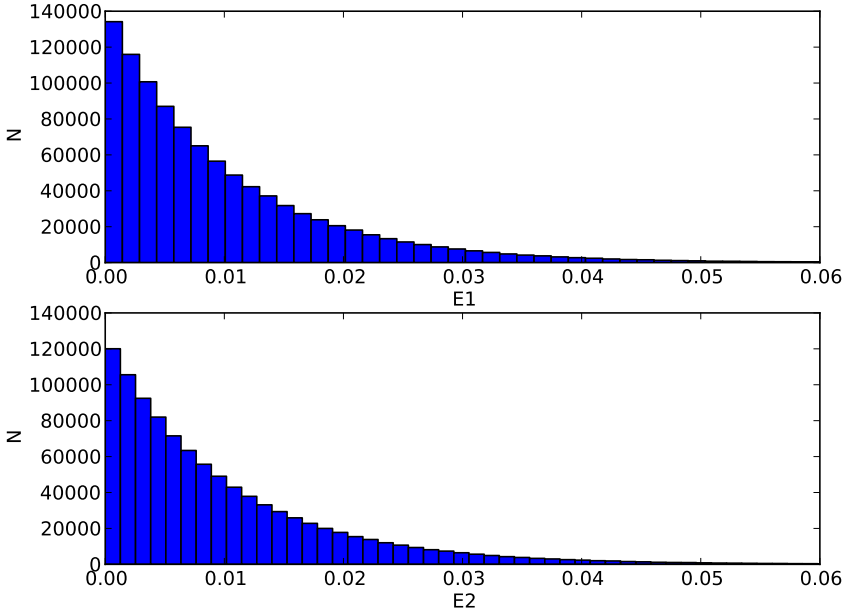
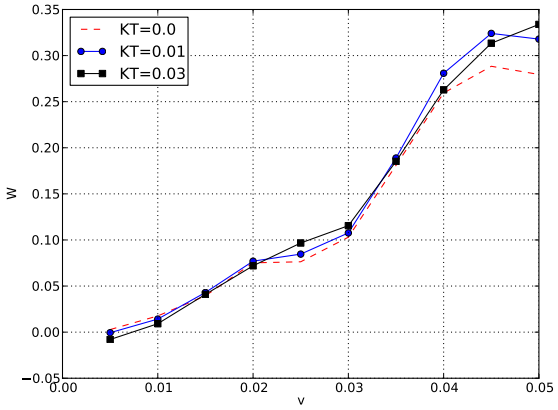


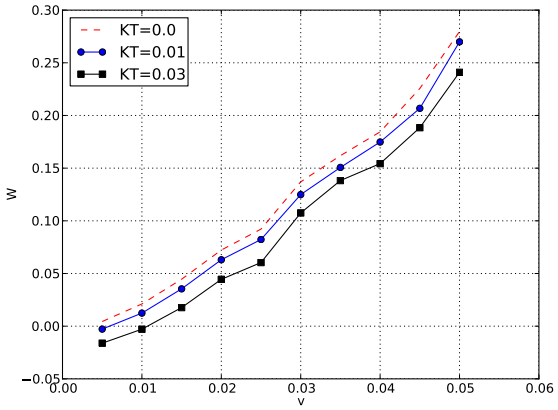
FIGURE 3.7: A plot showing the distribution of the initialised phonon energy for both the optical and the acoustic modes. E_1 and E_2 mark the optical and acoustic phonon modes, respectively. The distribution corresponds to the activation energy of $k_BT = 0.01\ \epsilon$

To explore the influence on the bond breaking events caused by excited phonon vibration modes at different temperature T, phonon vibrations are incorporated in calculation together with uniform sampling in the vibrating phase. The phonon energies used to initialise the phonon contribution were generated following the Maxwell distribution at given temperature T. As one example for the sampled phonon energies, Fig.3.7 shows their distribution at $k_BT = 0.01\ \epsilon$, both for the optical E_{opt} (upper panel) and acoustic modes E_{aco} (lower panel). Two independent calculations were performed for the activation energies of $T = 0.01\ \epsilon/k_B$ and $T = 0.03\ \epsilon/k_B$ respectively, with k_B the Boltzmann constant.

Fig.3.8 plots the consumed work W to break the bonding system against different separation velocities and panel (a) illustrates that for the near isotropic bonding case at the range of low separation speed or in other words, low loading rate, we can see that hot phonons helps the bond breaking to take place which corresponds to the thermo-activated regime in Fig.3.5. At $\mathbf{v} = 0.005\ \mathbf{v}_s$, a spontaneous breaking is even found with negative work consumed for the case of taking account of phonon activations. However,



(a)



(b)

FIGURE 3.8: The temperature contributions to the separation velocity (in units of the sound speed $\mathbf{v}_s$). In the figure plots the work W at different temperatures, which were incorporated in the calculation by the sampled phonon energies. Different bonding system was also tested including isotropic (a) and triple back bond system (b).

high temperatures act in the opposite direction as a resistance role when it comes to the larger separation speed or stronger external loading, in which case, the probability of breaking the back bond becomes overtaking that happens to the central bond and this make the bond breaking process approaches the catastrophic regime described in Fig.3.5. A crossover at $\mathbf{v} = 0.02\ \mathbf{v}_s$ is found consistent with the GAP/SW simulation [74]. Note that another crossover at $\mathbf{v} = 0.035\ \mathbf{v}_s$ between $T = 0.01\ \epsilon/k_B$ and $T = 0.03\ \epsilon/k_B$ may be attributed to the fact that larger convergence error is associated with the larger separation speeds $\mathbf{v}$ in the sampling of the phonon energies (see Fig.3.9). For the stronger back bond case [panel (b) in Fig.3.8], the central bond tends to break before that in the back bonds across investigated range of separation velocity. We find that temperature

contributes a prominent positive contribution for the overall bond-breaking process to take place and accordingly large propagation velocity of the crack tip if we move to take about the real crack system.

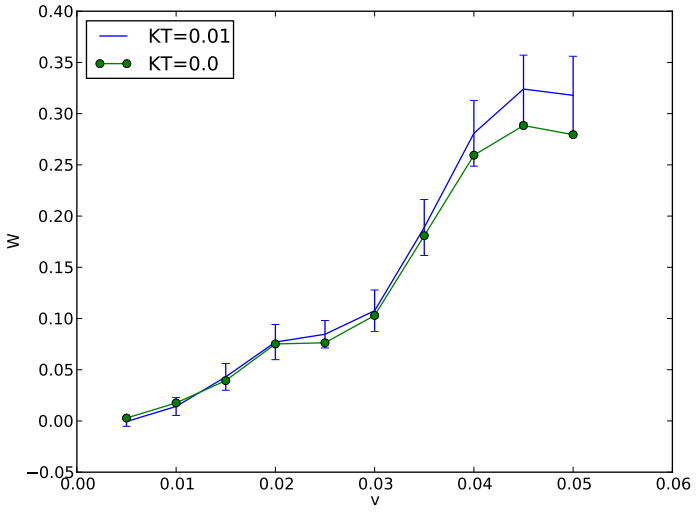


FIGURE 3.9: The plot showing the consumed work W vs separation velocity $\mathbf{v}$ with sampling variance included. Note that for quasi-static case (KT $= 0$), there is no sampling error and for KT $= 0.03$ there is larger variance than KT $= 0.01$.

3.5 Summary

In this section, the simulation of Silicon fracture was introduced and I also reviewed the work carried out with my collaborators during this thesis project which investigated the crack velocity dependence on temperature under the GAP/SW embedding scheme. In the case of no experimental observation available and chemical accuracy not assured, a theoretical model with LJ pair interaction was further used to probe the bond breaking mechanism. Under this model, crossover of the temperature contribution to the crack speed was found for the isotropic bonding type, while for the system with stronger back bonding system, the temperature favours the overall breaking process throughout the investigated range of separation velocities. We conclude that the crossover with respect to different temperatures, can be a typical phenomenon that exists in a domain of systems with a relatively uniform bonding structure or a bonding structure prone to temperature.

Chapter 4

Background- II

4.1 Machine Learning and ML Potentials

4.1.1 Introduction

Machine learning (ML) algorithms have been developed with the advent of supercomputing capacity. Data analysis and pattern recognition has grown in importance to cope with the large volume of information produced in such calculations. ML belongs to the broad subject of artificial intelligence with many applications in daily life, for instance, in the area of drug design, weather forecast, monitoring of ocean environment, and robotics *etc* [76]. As an inter-disciplinary subject, ML is closely connected with the development of computer science and many other modern technologies.

In the following sections, I introduce some algorithms used in machine learning, with emphasis on those useful for functional inference. Also, recent developments for the atomic potentials based on machine learning of the underlying potential energy surface will be reviewed, such as Gaussian Approximation Potentials (GAP), neural-network (NN) potentials, etc.

4.1.2 Bayes Theorem

For two dependent events A and B, their joint probability is written as: $\mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A)$.

$$\underbrace{\mathbf{P}(A|B)}_{posterior} = \frac{\overbrace{\mathbf{P}(B|A)}^{likelihood}\overbrace{\mathbf{P}(A)}^{prior}}{\underbrace{\mathbf{P}(B)}_{evidence}} \tag{4.1}$$

Bayes' Theorem is an approach to derive the *posterior* probability: $P(B|A)$, based on the *prior* knowledge: $P(A)$, likelihood: $P(B|A)$ and evidence: $P(B)$, as expressed in Eq.4.1.

An example is given below to demonstrate the Bayesian probabilistic view of the *posterior* probability:

suppose there is a drug test which gives a 99 % positive result to drug takers and 99 % negative results to non drug takers, and we also know that 0.5 % of people are drug takers. The question is: after knowing that for one person, the test result is positive, what is the probability that the person is drug taker ? To apply the Bayes' theorem, we consider that events A: the test person is drug taker, B: test result is positive. $P(B) = (1 - 0.5\ \%) \times (1 - 99\ \%) + 0.5\ \% \times 99\ \% = 1.49\ \%, P(B|A) = 99\ \%$, the *prior* $P(A) = 0.5\ \%$. The probability that the person being drug taker is calculated as :

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{99\ \% \times 0.5\ \%}{1.49\ \%} = 33.2\ \% \tag{4.2}$$

We can see that *posterior* probability is much lower than that would be expected from the individual event probabilities, 0.5 % or 99 % in the prior. This is a good example of the sometimes counter-intuitive results of applying a Bayesian approach.

4.1.3 Gaussian Process Regression

Bayes' theorem provides a statistical explanation of GP functional inference (the term 'Regression' refers to the fitting a curve across the data points even in cases where intrinsic noise is involved). Bayes' theorem states that the *posterior* probability can be related to the *prior* knowledge or experimental observations via the Eq.4.1, where $\mathbf{t}_N$ represents the N measured observables, and t_{N+1} is the $(N+1)$th measurement. The

probability for the t_N measurements is given by:

$$P(t_{N+1}|\mathbf{t}_N) = \frac{P(\mathbf{t}_N|t_{N+1})P(t_{N+1})}{P(\mathbf{t}_N)} = \frac{P(t_{N+1}, \mathbf{t})}{P(\mathbf{t}_N)} \tag{4.3}$$

4.1.3.1 Gaussian Processes

Gaussian Processes (GP) has been a useful tool for scientists to perform non-parametric function inference for many years. Compared to the parametric approaches such as the least-squares fitting, it is advantageous in more flexible functional form and no empirical constraints on the amount of parameters [77].

Let us start by expanding function $\mathbf{y} = f(\mathbf{x})$ with respect to a basis set $\{R_{nh}\}$ which are defined as the following:

$$R_{nh} \equiv \phi_h(\mathbf{x}_n) \tag{4.4}$$

R_{nh} therefore indicates the h-th basis function centred on the variable $\mathbf{x}_n$, Accordingly, the function $\mathbf{y}$ is written as :

$$y_n \equiv \sum_h R_{nh}\omega_h \tag{4.5}$$

where $\{\omega_h\}$ corresponds to the weight parameters for each of the basis function. The *prior* distribution of $\mathbf{w} = \{\omega_h\}$ is taken to be Gaussian type with zero mean and variance of σ_ω:

$$P(\mathbf{w}) = \mathbf{Normal}(\mathbf{w}, \sigma_\omega^2 \cdot I) \tag{4.6}$$

Since $\mathbf{y}$ is the linear combination of $\mathbf{w}$, its covariance matrix is :

$$Q = \langle \mathbf{y}\mathbf{y}^T \rangle = \langle \mathbf{R}\omega\omega^T\mathbf{R}^T \rangle = \mathbf{R} \langle \omega\omega^T \rangle \mathbf{R}^T = \sigma_\omega^2 \mathbf{R}\mathbf{R}^T \tag{4.7}$$

where $\mathbf{R} = \{R_{nh}\}$ indicates the basis set. The *prior* distribution of function $\mathbf{y}$ can therefore be expressed using the covariance matrix Q:

$$P(\mathbf{y}) = \mathbf{Normal}(\mathbf{y}, 0, Q) \tag{4.8}$$

Introducing C as the noise-included covariance matrix: $C = Q + I \cdot \sigma^2_{error}$, we obtain the following probabilities for $\mathbf{t}_N$ and t_{N+1}:

$$\begin{cases} \textit{Prior} \text{ probability :} & P(\mathbf{t}_N) \propto \exp\left[-\frac{1}{2}\mathbf{t}_N C_N^{-1}\mathbf{t}_N^T\right] \\ \text{Joint probability :} & P(t_{N+1}, \mathbf{t}_N) \propto \exp\left\{-\frac{1}{2}[\mathbf{t}_N \ t_{N+1}]C_{N+1}^{-1}[\mathbf{t}_N \ t_{N+1}]^T\right\} \end{cases}$$

In the above probabilities, C_N is the covariance matrix built for the data set $\mathbf{t}_N$ while C_{N+1} also includes the covariance with $(N+1)$th measurement: t_{N+1}.

Inverting of C_{N+1}. To make use of the inverting result for the covariance matrix C_N, we can write the C_{N+1} in the partition form comprising C_N. In the following matrix, K

$$C_{N+1} \equiv \begin{array}{|c|c|} \hline C_N & K^T \\ \hline K & \kappa \\ \hline \end{array}$$

represents the covariance between the $(N+1)-$th configuration and the N configurations in the database and κ the covariance between the test configurations and itself. The different parts of C_N^{-1} in the partitioned form can be calculated as:

$$C_{N+1}^{-1} \equiv \begin{array}{|c|c|} \hline M & \mathbf{m}^T \\ \hline \mathbf{m} & m \\ \hline \end{array}$$

$$\begin{cases} m = (\kappa - K^T C_N^{-1} K)^{-1} \\ \mathbf{m} = -m C_N^{-1} K \\ M = C_N^{-1} + \frac{1}{m}\mathbf{m}\mathbf{m}^T \end{cases}$$

The inverting of covariance matrix C_N is a costly process for large databases with a typical cost of $O(N^3)$. To circumvent this problem, in this thesis project, I adopted a sorting/selecting algorithm while keeping constantly dynamical training of large database possible. The overall cost of the calculation can be scaled close to $\sim O(N)$, as to be discussed in detail later (see Chapter 6).

Based on the above disscusion, the *posterior* probability : $P(t_{N+1}|\mathbf{t}_N)$ using the Bayes' theorem is thus expressed as:

$$P(t_{N+1}|\mathbf{t}_N) = \frac{P(t_{N+1}, \mathbf{t}_N)}{P(\mathbf{t}_N)} \propto \exp\left[-\frac{(t_{N+1} - \hat{t}_{N+1})^2}{2\sigma^2_{\hat{t}_{N+1}}}\right] \tag{4.9}$$

where the predictive mean and variance are:

$$\hat{t}_{N+1} = K^T C_N^{-1} \mathbf{t}_N; \quad \sigma^2_{\hat{t}_{N+1}} = \kappa - K^T C_N^{-1} K \tag{4.10}$$

4.1.3.2 Covariance Matrix

The general form of the covariance between two random variables $x^{(m)}$ and $x^{(n)}$ is expressed in Eq. 4.11,

$$C_{mn} = C(x^{(m)}, x^{(n)}; \theta) + \delta_{mn}\mathcal{N}(x^{(n)}; \theta) \tag{4.11}$$

where θ refers to the hyperparameters and $\mathcal{N}$ is noise model which is varied for the case of input-dependent data noise and typically constant for the case of input-independent data noise [77].

The covariance can take different forms and the requirement for the covariance matrix constructed upon is that it should be positive definite. In practice, a covariance functional form reflecting the physical nature of the target machine-learning functional is preferable to enhance the prediction accuracy. For instance, for machine learning of periodic functions, the forms of $\sin x$ or $\cos x$ are usually adopted. Different covariance functions ever emerged historically and can be found in literature while new covariance form are also under research in the community [76–78]. Among them, one of the most commonly used covariance at I-dimensional database takes the following form,

$$C(x^{(m)}, x^{(n)}; \theta) = \theta \exp\left\{-\frac{1}{2}\sum_{i=1}^{I} \frac{\left(x_i^{(m)} - x_i^{(n)}\right)^2}{l_i^2}\right\} \tag{4.12}$$

In the above equation, $x_i^{(m)}$ and $x_i^{(n)}$ is the i-th component of the data $\mathbf{x^{(m)}}$ and $\mathbf{x^{(n)}}$, respectively, θ is the hyperparameter, and l_i is the normalisation factor for the i-th dimension. Before the application of this techniques, I introduce another example for

the covariance which was often used in the early GP inference, i.e. the Matérn covariance. In this covariance, there is a hyperparameter ν which is adjusted to the extend of the required flexibility. By using different value for ν, a covariance function with different orders of differential smoothness resulted, as shown in Fig.4.1.

$$C(\mathbf{x}_1, \mathbf{x}_2, l, \nu) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left[\frac{\sqrt{2\nu}}{l} |\mathbf{x}_1 - \mathbf{x}_2| \right]^{\nu} B_\nu \left(\frac{\sqrt{2\nu}}{l} |\mathbf{x}_1 - \mathbf{x}_2| \right) \tag{4.13}$$

where B_ν represents the Bessel function of second kind of order ν, and l the characteristic correlation length scale.

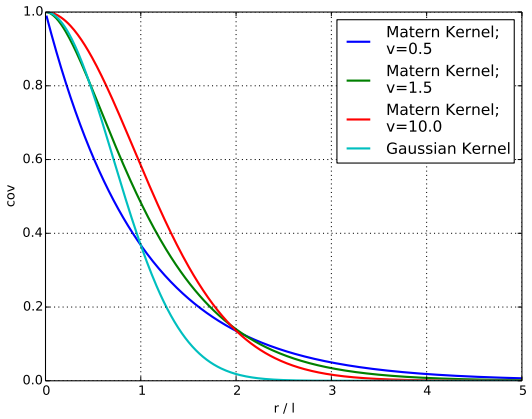


FIGURE 4.1: the covariance by definition of Matérn Covariance, in the plot $r = |\mathbf{x}_1 - \mathbf{x}_2|$. In the limit of $\nu \to \infty$, the Matérn covariance is equivalent to the Gaussian type which is infinitely differentiable.

- $\nu = 1/2$, $\exp\left(-\frac{r}{l}\right)$, it is equivalent to the Laplacian covariance, which typically governs the stochastic processes such as the Brownian motion.
- $\nu = 3/2$, $(1 + \frac{\sqrt{3}r}{l}) \exp\left(-\frac{\sqrt{3}r}{l}\right)$, once differentiable. Increasing ν, the covariance function becomes differentiable at higher order.
- $\nu \to \infty$, $\exp\left(-\frac{r^2}{l^2}\right)$, the function is infinitely differentiable, and equivalent to the Gaussian covariance.

Various covariance forms have been used by researchers in the area to meet their different needs in doing function inference, for instance, the Kriging method, and Radial covariance, etc. in the early stage of Gaussian Processes work [78]. In this thesis, the Gaussian processes on QM forces are performed with the following covariance form

(Eq.4.14), where d_{mn} is the distance between two atomic environments m and n, and σ_{cov} and σ_{error} are two hyperparameters (see section 4.1.3.3).

$$C_{mn} = \exp\left(-\frac{d_{mn}^2}{2\sigma_{\text{cov}}^2}\right) + \delta_{mn}\sigma_{\text{error}}^2 \tag{4.14}$$

4.1.3.3 Hyper-parameters

The parameters that are present in the kernel of the covariance are called hyperparameters, which is used to tune the predicted function form or regularise the function inference process. Two hyperparameters $\{\sigma_{\text{cov}}, \sigma_{\text{error}}\}$ are involved in the covariance (Eq.4.14)

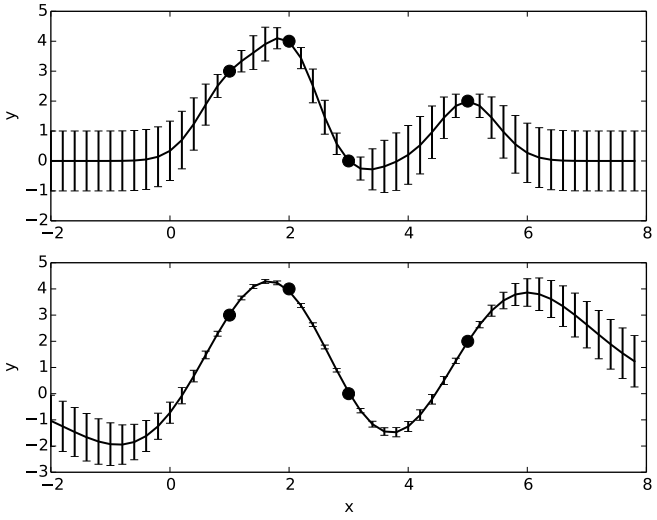


FIGURE 4.2: One-dimensional Gaussian Process with the black dots mark the data points. The solid curves marks the regressed mean function with the error bars indicating the prediction variance from the GP process. σ_{err} =0.05 eV/Å and two different σ_{cov}= 0.5 (upper) and1.5 (lower) respectively in the plot.

used in this thesis. A demonstration of the inference for a one-dimensional function is given in Figs.4.2 and 4.3, with several different *prior* hyperparameters $\{\sigma_{\text{error}}, \sigma_{\text{cov}}\}$. The σ_{cov} is significant in controlling the correlation length of any two data points along the distance scale. Smaller σ_{cov} makes the prediction more accurate locally near the data points, but increases the uncertainty in the longer extrapolation regime, in other words, decreases the weight of distant data points in the prediction. The larger σ_{cov} tends to put less weight on the local data points and usually yields a more general predictive form in the long range of data space.

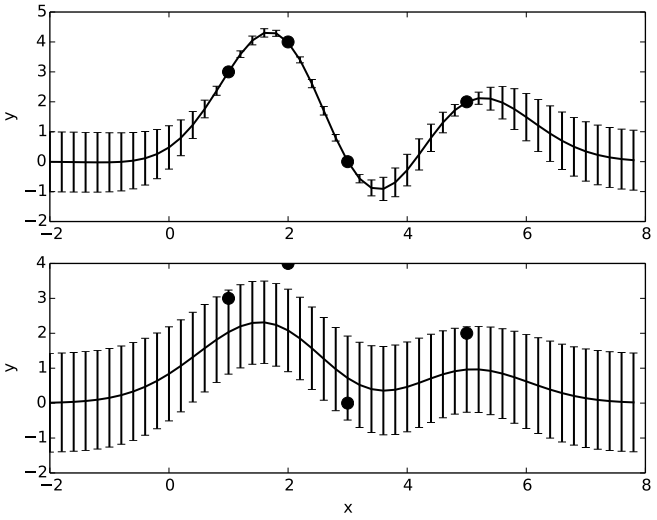


FIGURE 4.3: Two different noise assumed for the Gaussian processes, σ_{error} of 0.05 (upper) and 1.0 (lower) eV/Å. The other hyperparameter σ_{cov} were kept constant for the two cases as 1.0

The noise assumed on the data is indicated by σ_{error}. Small values correspond to the rigid parameterisation with over fitting to the data points, which however, lack of the extrapolation capability beyond the data points. Larger σ_{error} enables better extrapolations at a risk of losing the accuracy at the fitting data (with 'blurring' around the data of the magnitude of σ_{error}) and predicted function takes simple form rather than cross the accurate data points. In the case of high-accuracy required calculations, suitable hyperparameters are key and thus optimisation procedures are often needed, such as by the maximising of the marginal likelihood to be discussed below.

As an example for 2-dimensional GP prediction, I adopted the Euclidean distance between the variable vectors to construct the covariance matrix (Eq.4.14) and the predicted mean function is found in Fig.4.4. The mean prediction surface is seen to be a product of the multi-variant 2D Gaussians centred on each data points and zeros values are found where there is few data distributed. Extending to even higher-dimensional case, the database is re-organised by the adopted covariance form and the function regression is based on an inversion relation with respect to this constructed data topology. The prediction becomes enormously challenging as more complexities are incorporated.

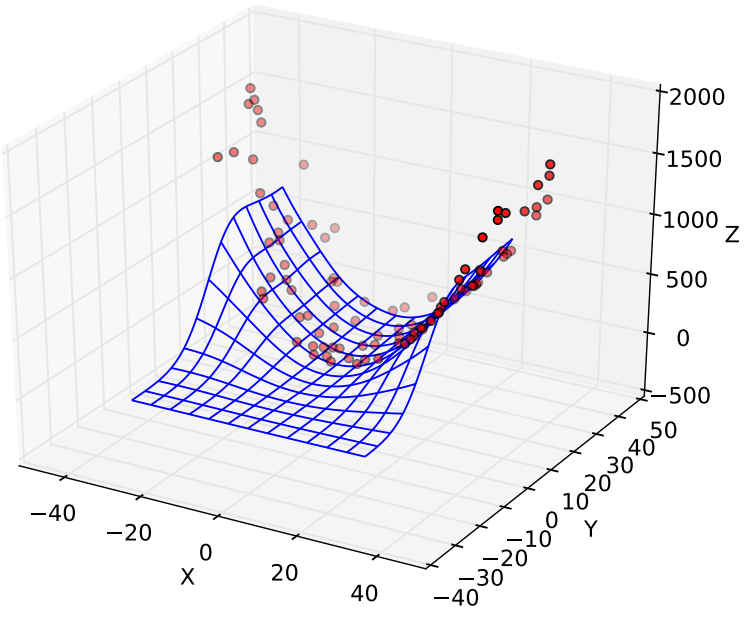


FIGURE 4.4: Example plot showing the function inference in 2D data space. The function surface (blue) was inferred from 100 noisy data points (red solid dots). The covariance matrix was constructed using the Euclidean metric. Hyper-parameters in this Figure are: $\sigma_{cov} = 10.$ and $\sigma_{error} = 0.05$

4.1.3.4 Hyper-parameter optimisation

The optimal hyperparameters for a given learning set can be found by maximising the logarithm of the likelihood with *prior* hyperparamters: $L(\theta) = \log(P(\mathbf{t}|\mathbf{x}, \theta)); \theta \equiv \{\sigma_{\text{err}}, \sigma_{\text{cov}}\}$ [76].

$$L = \log(P(\mathbf{t}|\mathbf{x}, \theta)) = -\frac{1}{2}\mathbf{t}^T C_N^{-1}\mathbf{t} - \frac{1}{2}\log|C_N| - \frac{N}{2}\log(2\pi) \tag{4.15}$$

The likelihood is comprised of two contributions, i.e., the *data fitting* term: $-\frac{1}{2}\mathbf{t}^T C_N^{-1}\mathbf{t}$ and *complexity penalty* term: $\frac{1}{2}\log|C_N|$. The third term: $\frac{N}{2}\log(2\pi)$ is associated with dimension of the database and is constant for a fixed database. The optimal Gaussian Process inference tends to be a balance between these two considerations.

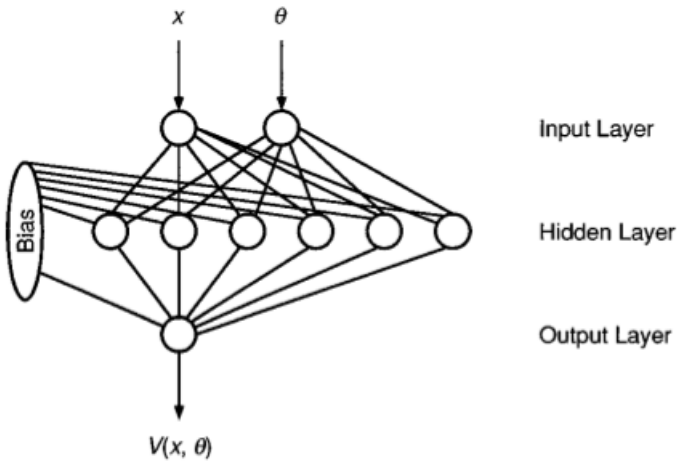


FIGURE 4.5: A schematic plot showing the NN prediction with two-dimensional input variables (x, θ) and bias in the hidden layer [79].

To maximise L, the derivative of L with respect to the hyperparameter θ is expressed as following and optimal hyperparameters are thus derived by numerically solving $\frac{\partial L}{\partial \theta_i} = 0$:

$$\frac{\partial L}{\partial \theta_i} = \frac{1}{2}\mathbf{t}^T C_N^{-1} \frac{\partial C_N}{\partial \theta_i} C_N^{-1}\mathbf{t} - \frac{1}{2}\text{trace}\left(C_N^{-1}\frac{\partial C_N}{\partial \theta_i}\right) \tag{4.16}$$

As can be seen from Eq.4.16, the computational cost of the above procedure can be high as the inverting of covariance matrix C_N with *prior* hyper parameters is involved and the optimisation usually has to be performed n runs for convergence. Different algorithms have been developed to cope with the training in large database, one of which is the sparisification algorithms [76].

4.1.4 NN algorithms

Neural Networks (NN) are another function inference algorithm inspired by biological neurones. The algorithms works by mapping the data from the 'input layer' onto the 'hidden layer' with parameters iteratively optimised and used to make predictions for new inputs. Under this algorithm, the input functions are mapped by the iterative optimised weight parameters in the 'hidden layer' and used to make prediction for other general input representing functions. Neural networks works by implementing a function y ($\mathbf{x}$, $\mathbf{w}$), and optimisation of the weight parameters space $\mathbf{w}$ and output y as non-linear

function of the input $\mathbf{x}$ space, e.g. signal function [77] as follows,

$$y(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp\left(-\sum_i w_i x_i\right)} \tag{4.17}$$

A schematic plot is given in Fig.4.5, prediction from two-variable inputs and a bias exerting on the hidden layer.

4.1.5 Summary

In this part, I have introduced the Gaussian Process function inference, starting from the point of view of Bayes' theorem concerning *posterior* probability distribution. Though the Gaussian Processes are far from new algorithm and have been widely used by scientists for decades, they however, increasingly attract attentions by the work that required with upgrading computational power. Applying the technique to material simulation however, we can see lots of potential while the simulation scale is proceeding to higher level and the gap between the computer calculation and real experiment observation is narrowing. Using function inference from ML to fit the potential energy surfaces has been adopted by several groups and resulted into a series of machine-learning potentials and attempts to address multi-scale problem was also made in practice. In the following section, I will review these potentials with their highlighted application in material simulations.

4.2 ML Potentials

4.2.1 Introduction

In the following, I will review the ML potentials which are typically free from explicit fitting parameters. Rather, they adopt function inference from ML techniques based on a QM database. By representing the configurations with suitable descriptors designed for QM energy learning, these potentials are used to predict the first-principles PES, atomisation energies or density functionals with respect to the electron density. The procedures taken by the implementation of the ML potentials are typically: (1) representing the configurations taking into account the associated symmetries with the energy quantity; (2) performing function inference on the PES or the atomisation energies using the ML algorithms, such as Gaussian Processes (GP), Neural Networks (NN), Kernel Methods; (3) hyperparameter optimisation.

4.2.2 Representation of the Atomic environments

Many representation schemes for the chemical environments are used in the characterisation of the atomic structure in molecular dynamics. The local bond-orientation order parameters were proposed by Steinhardt *et al.* where the rotational invariance is intrinsically captured by using the basis set of spherical harmonics. These parameters have been found to be generally useful in discriminating the structure types in liquids, glasses and solid materials [80].

Expanding the bond orientation $\{\hat{r}_{ij}\}$ with spherical harmonics Y_{lm}, Q_{lm} is defined as,

$$Q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\hat{r}_{ij}) \tag{4.18}$$

In Eq.4.18, $N_b(i)$ represents the number of nearest neighbours of the i-th atom, and $Y_{lm}(\hat{r}_{ij})$ corresponds to the spherical harmonic for $\hat{r}$ under index of m and l and $|m| < l$. The index i indicates the i-th atomic configuration and j runs over the neighbouring atoms.

The Steinhardt order parameters are rotationally invariant and constructed as,

$$Q_l(i) = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{m=l}|Q_{lm}(i)|^2}. \tag{4.19}$$

and

$$W_l(i) = \sum_{m_1,m_2,m_3}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \times Q_{lm_1}Q_{lm_2}Q_{lm_3}$$

Modified bond-order parameters were proposed recently based on averaged local bond order parameters as given in Eq.4.20 and shows very good resolution in determining crystalline structures [81],

$$\bar{Q}_l = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{m=l}|\bar{Q}_{lm}(i)|^2}. \tag{4.20}$$

and

$$\bar{W}_l(i) = \sum_{m_1,m_2,m_3}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \times \bar{Q}_{lm_1}\bar{Q}_{lm_2}\bar{Q}_{lm_3}$$

In the above, $\bar{Q}_{lm} = \langle Q_{lm}\rangle$ gives the average taken over the set of bonds for the neighbouring atoms and the coefficients

$$\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix}$$

in the third-order invariants $W_l(i)$ are Wigner $3j$ symbols and produce zero unless $m_1 + m_2 + m_3 = 0$.

A useful combination descriptor which is sensitive to structure difference and orientational symmetries is,

$$\hat{W} \equiv W_l \Big/ \left(\sum_{m=-l}^{m=l}|\bar{Q}_{lm}(i)|^2\right)^{3/2}$$

The above descriptors are not complete representations, which means that for wider spectra of structural variation, two different atomic environments may fall into the same representation. A complete descriptor was developed by Bartok *et al.* based on the bispectrum of the atomic environments and this descriptor was applied into their

machine learning potentials (GAP, see next section) in the early stage [9], though it was found problematic in convergence with respect to the number of neighbouring atoms within atomic environment. The recent proposed smooth-overlap-of-atomic-positions (SOAP) descriptor combining the bispectrum and overlap of atomic position calculation has improved performance [82]. Other descriptors were developed for the purposes of the different ML schemes, of which I will give a review below.

4.2.3 Gaussian Approximation Potentials

One of the ML potentials is the Gaussian Approximation Potential (GAP) developed by Bartok and his collaborators [9] and they are specifically implemented using the Gaussian process inference of the PES surface. In the scenario of a GAP, the total energy E_{total} is constructed to be sum of the atomic energy $\{\epsilon_i\}$ centred on each of the atoms and the function relation between atomic energy and atomic environments are the underlying target for the ML techniques to address. For localised bonding, such as covalent bonds, this local energy concept is justified. For long-range interactions, such as the Coulomb or dispersion, these terms are usually separated from the local energies $\{\epsilon_i\}$ to assure a much better localisation for the ML to perform [9].

$$E_{\text{total}} = \sum_{i=1}^{N_{atoms}} \epsilon_i(\mathbf{x}_i) + \sum_{ij} \frac{q_{ij}}{r_{ij}}. \tag{4.21}$$

In the GAP potentials, the predicted atomic energy for new configurations $\mathbf{x}_i$ are expressed as,

$$\epsilon_i(\mathbf{x}_i) = \sum_{n=1}^{N_{data}} \alpha_n k(\mathbf{x}_i, \mathbf{x}_n) \tag{4.22}$$

where $k(\mathbf{x}_i, \mathbf{x}_n)$ indicates the covariance element between the i-th test configuration and the n-th reference configuration in the QM database. $\{\mathbf{x}_n\}$ and α_n are product of inverted covariance matrix C^{-1} and the QM energy entries $\mathbf{y}$ of the database,

$$\{\alpha\} \equiv \alpha = C^{-1}\mathbf{y} \tag{4.23}$$

In GAP, the learning of QM forces and/or Viral stress are technically incorporated by the derivative of the covariance with respect to the atomic coordinates, while the force prediction for each atomic configuration is carried out as the derivative of the predicted

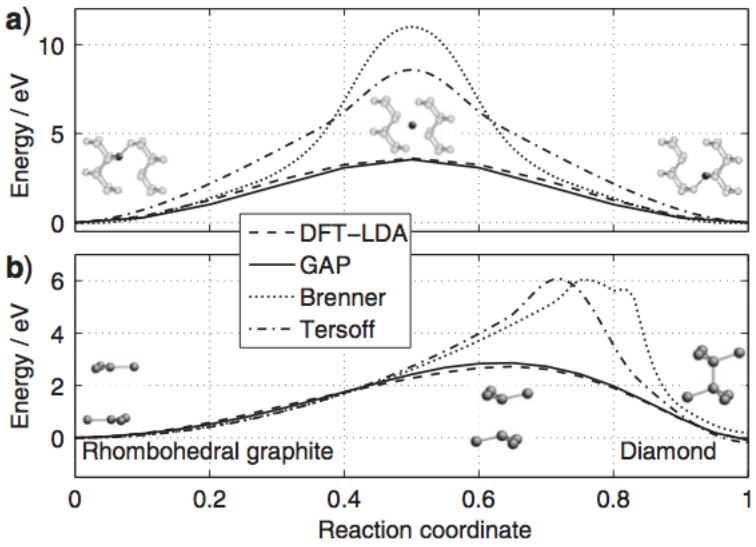


FIGURE 4.6: The energetics was calculated using classical force fields, GAP potential and DFT for (**a**) the linear transition path and (**b**) the structural transformation from rhombohedral graphite to diamond-type carbon. The figure was reproduced from[9]

energy quantity. In Fig.4.6, GAP potentials were used to calculate the structural transition path of Carbon. Accuracy comparable to the first-principles level was achieved with a database which was generated from an MD trajectory [9]. GAP potentials have also been used into the crack simulation of Silicon (see section 3.3) and the simulation of molecular, condensed water and tungsten [83, 84].

4.2.4 Neural Network Potentials

Based on the Neural Network (NN) algorithm (section 4.1.4), another machine-learning potential scheme was ever established by Behler and Parrinello in 2007 [10]. In NN potentials, energies are taken as the sum of the atomic energy which functionally depends on the local atomic configurations. The atomic configurations however, are represented by introducing a symmetry function of the neighbouring atomic positions. The NN prediction of PES is further performed at the approximation of high-dimensional representation [10].

The symmetry functions includes the two-body and three-body contributions, which are sum of Gaussians with controlling parameters of η and R_s and cutoff function $f_c(R_{ij})$.

Explicitly, the two-body symmetric functions are given by,

$$G^2 = \sum_{j \neq i}^{all} \exp\left[-\eta(R_{ij} - R_s)^2\right] f_c(R_{ij}) \tag{4.24}$$

while the symmetry function from three-body contributions is:

$$G^3 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos\theta_{jk})^{\zeta} \exp\left[-\eta(R_{ik}^2 + R_{jk}^2 + R_{ij}^2)\right] f_c(R_{ik}) f_c(R_{jk}) f_c(R_{jk}) \tag{4.25}$$

where θ_{jk} indicates $\arccos\left(\frac{\vec{R}_{ij} \cdot \vec{R}_{ik}}{R_{ij} \cdot R_{ik}}\right)$. The cutoff function $f_c(R_{ij})$ takes the decreasing form of $\left[\cos\left(\frac{\pi R_{ij}}{R_{cut}}\right) + 1\right]/2$ within the neighbour cutoff radius R_{cut}, and are set to be zeros for $R_{ij} \geq R_{cut}$,

$$f_c(R_{ij}) = \begin{cases} \left[\cos\left(\frac{\pi R_{ij}}{R_{cut}}\right) + 1\right]/2 & \text{for} \quad R_{ij} < R_{cut}, \\ 0 & \text{for} \quad R_{ij} \geq R_{cut} \end{cases}$$

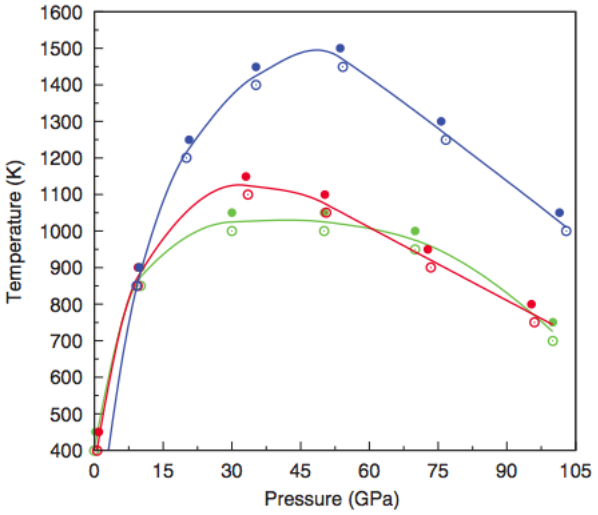


FIGURE 4.7: The melting curve of Sodium subject to external pressures. Comparison was shown for the NN potential (green), an effective pair potential based on jellium model or uniform electron gas model (red), and repulsive wall of effective pair potential (blue). At the pressure of 90 Gigapascal (GPa), a structural transition from (body-centered-cubic) *bcc* to (face-centered-cubic) *fcc* was taken into account in the calculated curves. The figure was reproduced from literature [85].

The NN potentials for example have been used to simulate the melting of Sodium and exploration of the nucleation mechanism of graphite / diamond transition [85–87]. As shown in Fig.4.7, where the abnormal melting behaviour was displayed in the study of

the melting process in the Sodium system under pressure. Other application includes into the zinc oxide and water [88, 89].

4.2.5 ML Model for Atomisation Energy

Machine learning of the molecular energy quantities is of significant research interest and challenge. In the scheme developed by Rupp et al., one of the key parts, the representation for the molecular configurations, was constructed by a Coulomb matrix, whose elements M_{IJ} correspond to the atomic energy and Coulomb interaction energy as expressed in the equation below,

$$M_{IJ} = \begin{cases} 0.5Z_I^{2.4} & \text{for } I = J \\ \frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|} & \text{for } I \neq J \end{cases}$$

where the $\{Z\}$ and $\{R\}$ are the nuclear charge number and atomic positions, respectively.

ML performance of this scheme for organic molecular was systematically demonstrated by calculations upon a database containing molecular configurations that are stable by the knowledge of the criteria in organic chemistry and also accessible to the synthetically experiments [90].

4.2.6 ML of Electron Density Functionals

Apart from the machine learning of the energy quantities, I note about function inference on the electron density functionals. Snyder et al. made ML predictions in the prototype case of 1-dimensional non-interacting spinless fermions and the learning of kinetic energy functional $T(\mathbf{n})$ was achieved within chemical accuracy of 1 kcal/mol (or 0.043 eV) [91]. To calculate the functional derivative $\nabla_{\mathbf{n}} T(\mathbf{n})$ in the density space $(\mathbf{n}_1, \cdots, \mathbf{n}_j)$, principal component analysis (PCA) is carried out on m relevant reference density to find the l density dimensions with the largest variation with respect to the predicting density $\mathbf{n}$. A comparison of the ML approximation and exact self-consistent result is given in Fig.4.8, where $\mathbf{P}_{m,l}$ indicates the density projection matrix onto the l principal density dimensions. Based on the derivative (or gradient) prediction, starting from a guessed density, optimisation procedure can thus be performed to find the density that corresponds to the minimal total energy. The accuracy of this derivative calculation

however, is limited and the yielded density from the minimisation may be different, depending on the initial guess.

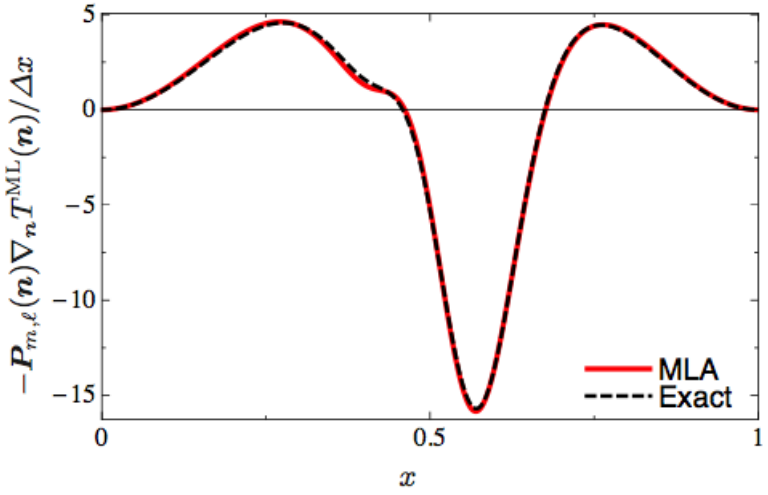


FIGURE 4.8: Projected kinetic energy functional derivative for one-dimensional non-interacting fermions is plotted for both the ML approximation (MLA) and the exact self-consistent result. Figure is from [91].

4.2.7 Summary

In this chapter, machine learning algorithms to carry out function inference were introduced and the schemes for ML potentials and density functional were reviewed. Providing the database containing QM energies, the potential energy surface is functionally regressed at a high accuracy. The applications and developments were also briefly introduced in the materials simulation process. The limitation of these ML potentials is that, they still work like classical potentials (e.g. energy conserved, validation of database required, transferability limited) after the QM data training and their transferability to different atomic environments largely relies on the completeness of the training database. In the following Chapter, based on a philosophy different from the static ML potentials while targeting the practical computational efficiency and applicability to large-scale MD simulations, an accurate machine-learning scheme for QM force calculation will be presented.

Chapter 5

Results II: Machine Learning of QM Forces

5.1 Motivation for ML of QM Forces

ML potentials have been introduced in the previous chapter and it was shown that they essentially work like classical potentials after teaching with QM database. Accordingly, during MD simulations, configurations which turn out to be beyond the knowledge represented within the teaching database will be predicted with much less reliability. To address this problem, instead of adopting a 'once-and-for-all' learning methodology, we will construct a dynamic learning approach. The database in this new learning approach is dynamically updated when and only when novel configurations beyond the reliable regime of the Bayesian inference prediction are encountered. With a database growing when chemical novelty is encountered, the predicted atomic forces will be typically bounded closer to the first-principles target. Significantly, the prediction variance which comes naturally along with the ML procedure will also be made use of to regulate the QM-database augmentation during dynamics. This machine learning force calculations work 'on-the-fly', for which, we denote as MLOTF.

The prediction of force by analytical differentiation of the predicted total energy E results in much amplified uncertainty. A plot illustrating this statement is given in Fig.5.1 including the error comparison between predicted energy and its differential product, atomic forces. From the plot, we can see that, large force errors are also present

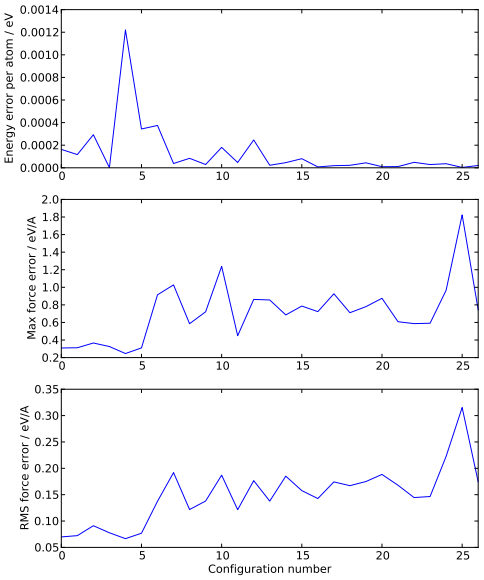


FIGURE 5.1: In the plot, the energies were calculated under the machine learning potential of GAP model, and forces by analytical differentiation upon the ML energies. From top to bottom, the plots give the energy error, maximum (Max) and root-mean-square (RMS) force error against the corresponding first-principles results.

where the predicted energy is over precise, for instance, the 25-th atomic configuration has a prediction energy error ΔE close to 0, while the corresponding force error is higher than 0.25 eV/Å. In this case, much more QM data for the configurations deformed from the 25-th configuration would be required for an accurate description of the forces.

MLOTF methodology to be discussed below, is an approach aiming for high-precision force prediction by directly machine learning from the QM force database, without invocation of the energy quantities. The target learning function is thus not constrained to the PES. Instead, particular emphasis is put onto the force, or the gradient from the PES. Significantly under this new scheme, the implementation into large-scale MD simulation can be elevated by the on-the-fly machine learning.

5.2 Possibility for ML of QM Force

In most covalent materials, quantum mechanical force exerted on atom depend on the local atomic environment formed by the neighbouring atoms. Such QM force can generally be approximated within a certain precision by the calculation upon a finite cluster

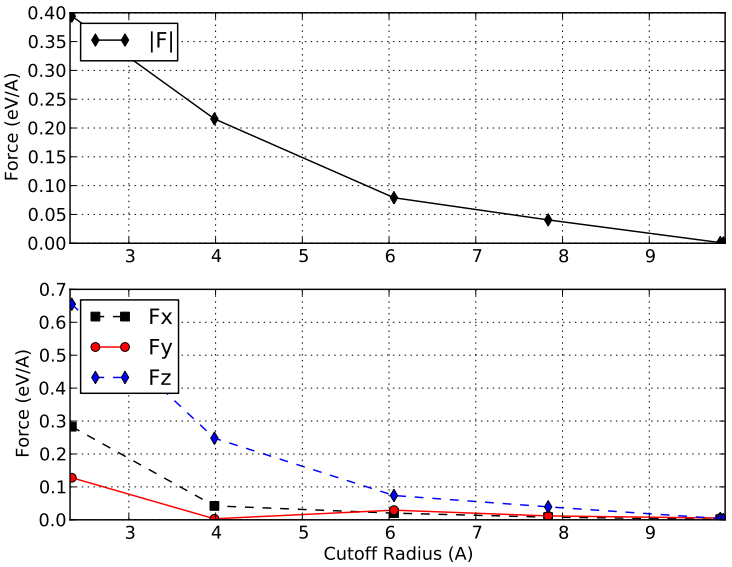
with a cutoff number of neighbouring atoms. As displayed in panel (a) of Fig.5.2 for Silicon, for local cluster calculation around a centre atom with neighbouring atoms cut-off at above 8 Å in spherical radius (around 4 bond hops), the DFTB forces converge within a precision of 0.05 eV/Å in magnitude, or less than 4% in relative error. At 10 Å or 5 bond hops, the relative error decreases to 1% or $\sim$ 0.01 eV/Å in magnitude. In Panel (b) the force convergence is tested on more general structural forms. Similarly, for cluster above cutoff radius of 8 Å, forces convergence with error less than 0.05 eV/Å. For the MLOTF in this thesis work, we will typically adopt a cluster cutoff at 8 Å for Si calculations, while a Gaussian noise $\sigma_{\text{error}} = 0.05$ eV/Å is assumed for the QM force data. Force convergence at a cutoff cluster size suggests it is computationally robust to derive a scheme only taking into account the atomic environments pertinent to the QM force. Especially in the non-periodic large system calculations, such kind of cluster calculation within a reasonably cutoff radius is usually used instead of doing the self-consistent calculation treating the system as a whole [12, 13].

Therefore, ML of QM forces can be performed by function inference from its relation with the local atomic environments. To this end, it is desirable to derive a descriptor to account for the features of the local atomic environment as completely as possible to achieve the best accuracy. As is well known, the higher the dimensionality incorporated in the simulation, the more computationally costly it becomes. Also the famous 'curse' of high dimensionality in data learning makes it worthwhile effort to derive a pertinent and dimension-reduced atomic description [93]. This forms the topic for the following part.

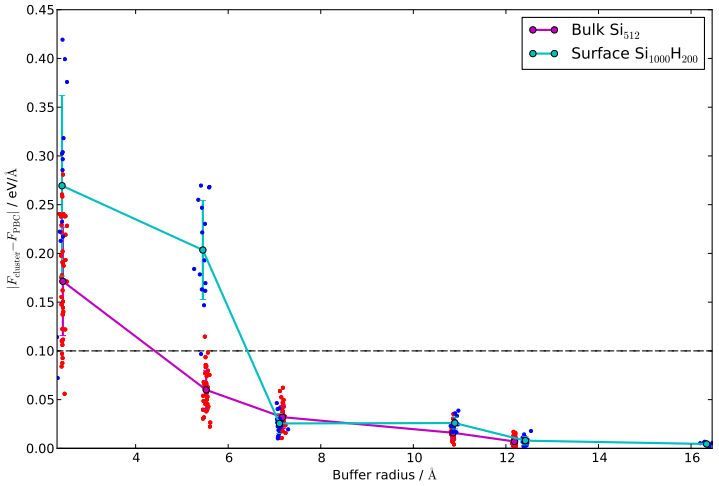
5.3 Representation for the Atomic Environments

The necessity of developing representations for atomic environments has been discussed in Chapter 4. To our knowledge, there is no available representation scheme for doing machine learning on forces, in which case, typical descriptor developed for ML of energy quantity does not work.

The local atomic environment associated with the QM force embraces the $SO(3)$ symmetry group. However, if described in the usual Cartesian coordinate, the orientation



(a)



(b)

FIGURE 5.2: Convergence of DFTB forces in Silicon systems with respect to the atomic cluster size. (a) convergence of the force magnitude and its Cartesian components. The forces converge within ~ 0.05 eV/Å at a cutoff radius of 8.0 Å. The test configuration comes from a typical MD trajectory run at 1000 K using DFTB Hamiltonian. (b) shows the convergence of the DFT forces on the configurations sampled for bulk Si_{512} and Si surface terminated with Hydrogens ($Si_{1000}H_{200}$). In comparison with the accurate PBC calculation, the forces converge within the precision 0.05 eV/Å for cutoff radius above 6 Å. The lower plot was reproduced from Reference[92].

of force depends on the specific choice of the reference frame. The more general the information can be represented, the better accuracy and efficiency can be achieved in the prediction calculations. Intrinsic symmetries like rotational, reflection, inversion, and

permutation should be incorporated into the representation for atomic environments. Apart from the symmetry reducibility, the representation should be complete in terms of capturing features of the atomic environments associated with the QM force, as well as having a smooth relationship with respect to variations in the atomic positions. As one of the key results, I will describe two schemes that were developed during this thesis work: (1) the overlapping measurement of atomic environments (2) internal-vector representation.

5.3.1 Distance by Overlapping Measurement

This representation scheme works by distance mapping according to the overlapping measurement of neighbouring atoms and was developed with my collaborators at King's College London. In this section, I present force calculation results with this method. In this representation, the position of the neighbouring atoms $\{\vec{r}_i\}$ are placed by centred delta functions, an atomic density function ρ is the sum of these delta functions with a cutoff function f_{cut}:

$$\rho(\vec{r}_0) = \delta(\vec{r}_0) + \sum_i \delta(\vec{r}_i - \vec{r}_0) f_{cut}(|\vec{r}_i - \vec{r}_0|), \tag{5.1}$$

where $\vec{r}_0$ indicates the centre atom and can be shifted to the origin when comparing two atomic density functions. The $f_{cut}(r)$ corresponds to the cutoff function and a form is given below

$$f_{cut}(r) = \begin{cases} 0 & r > r_{cut} \\ \frac{1}{2}\left[\cos\frac{\pi(r - r_{cut} + r_{tran})}{r_{tran}} + 1.\right] & r_{cut} - r_{tran} \leq r \leq r_{cut} \\ 1 & r < r_{cut} - r_{tran} \end{cases}$$

The meaning of the above cutoff function lies in that only the neighbouring atoms within the cutoff r_{cut} are taken into the overlap measurement between the atomic configurations. The introducing of the parameter r_{tran} ensures the smooth transition associate with the atomic movements across the cutoff radius r_{cut}.

The overall distance is constructed by integration of the overlapping measurement of two sets of Gaussians in the rotational space, as expressed in Eq.5.2:

$$d^2(\rho_1, \rho_2) = \int |\rho_1(\vec{r}) - \rho_2(\vec{r})|^2 dr^3 \tag{5.2}$$

To combine the rotationally-equivalent images in one representation, the distance is always minimised with respect to the rotation on one of the configurations to achieve the minimum distance.

$$D_{1,2} = \min_{\hat{R}} d^2(\rho_1, \hat{R}\rho_2), \tag{5.3}$$

where $\hat{R}$ indicates the rotational operator. Accordingly, the covariance is constructed as:

$$\mathrm{cov}(\rho_1, \rho_2) = \theta \cdot \exp\left[-\frac{D_{1,2}}{2\sigma^2_{\mathrm{cov}}}\right], \tag{5.4}$$

where θ is the normalisation factor.

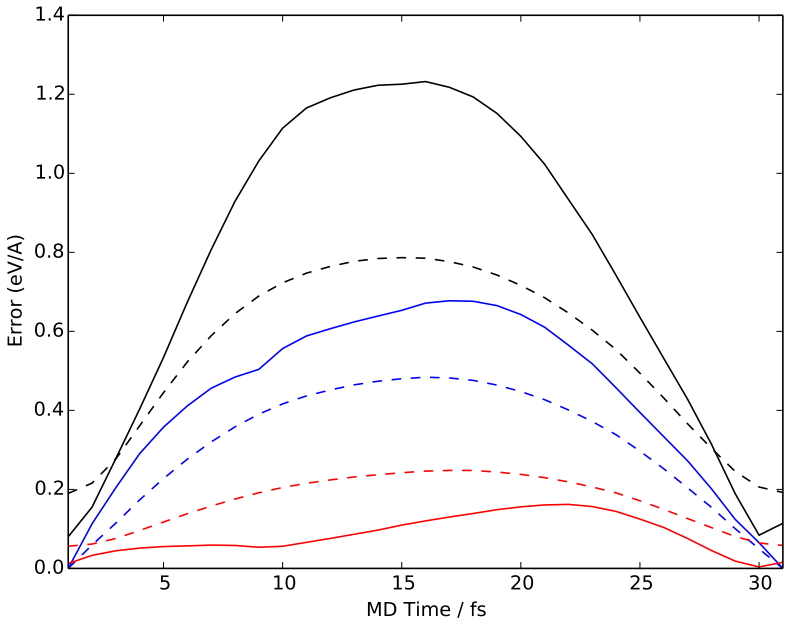


FIGURE 5.3: The density overlap distance (Eq.5.3) was used for constructing the covariance matrix in the Gaussian Process prediction of QM forces. The accuracy test was performed on MD trajectory of bulk Silicon at 1000 K in the predictor/corrector way. Different number of teaching configurations: 10 (black), 20 (blue), and 50 (red) were used from the past trajectory with time interval of 30 fs. However, it is noted that alignment was not problematic in this plot because of the limited time range of the trajectory.

Based on the distance in Eq.5.2, force machine learning was tested on configurations along the bulk MD trajectory (Fig.5.3) and good results were obtained for a relatively

small teaching database. For large database, where rotation becomes frequent issue, the method may not work due to the fact that it involves numerically minimising the overlap distance D_{12} with respect to all the rotational images. The rotations are numerically represented in the 4-dimensional quaternions space [94] and distance in Eq.5.2 is calculated by searching for the global minimum. This minimisation process becomes a bottle neck in computation and can be stuck in the local minima, in analogy to that in the structure-searching research [95]. In the following, I will introduce an information-efficient approach which represents the system with symmetrically-reduced internal vectors.

5.3.2 Internal Vector Representation

As illustrated in Fig.5.4, instead of describing the system with external Cartesian coordinates which are made up of $\{\vec{U}_i\}$ (i=1, 2, and 3), we can derive a set of internal vectors $\{\vec{V}_i\}(i = 1, 2, \cdots, \text{and } k)$ following the same symmetries as the QM force and the atomic environment. A symmetrically-reduced representation can be further constructed using these internal vectors, explicitly by describing all the vectors in this coordinate system comprised of k vectors. For $k > 3$, this typically forms an over-determined coordinate system. Machine learning techniques, e.g. GP, can be further applied on the predictions of the force components on each of the internal directions.

To satisfy the symmetry requirement, the internal vectors can take simple form as the linear sum of the bond direction vectors in the real space. We further smoothly screen the interactions from neighbouring atoms above the cutoff inter-atomic distance: $r_{ij} \geq r_{\text{cut}}$ (Eq.5.5) by using a radially-dependent weight factor $\omega_i = e^{-\left(\frac{r_i}{r_0}\right)^m}$ which reflects the decaying contribution from the neighbouring atoms with respect to the increased distance r_i. Since the weights $\{\omega_i\}(i = 1, \cdots, N_{\text{neighb.}})$ only involve the magnitude of the vectors, it can be demonstrated that these internal vectors intrinsically satisfy the same symmetries as the target QM force.

$$\vec{V} = \sum_{i=1}^{N_{\text{neighb.}}} \hat{\mathbf{r}}_i \cdot w_i = \sum_{i=1}^{N_{\text{neighb.}}} \hat{\mathbf{r}}_i \cdot e^{-\left(\frac{r_i}{r_0}\right)^m} \tag{5.5}$$

By varying the parameters r_0 and m in the weight function, a set of internal vector can be derived from Eq.5.5. These vectors have two notable features: (i) they share the

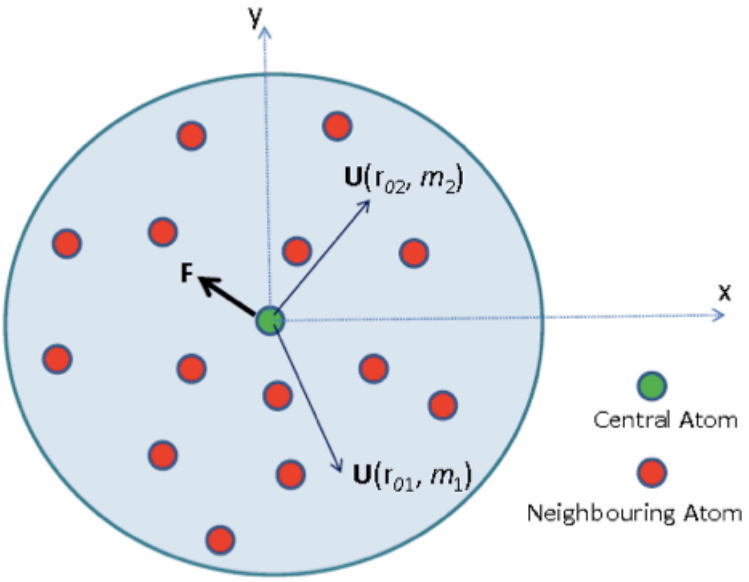


FIGURE 5.4: Two dimensional schematic plot showing an atomic environment within a spherical cutoff radius r_{cut}. Both external coordinate $(\vec{x}, \vec{y})$ and internal coordinate $(\vec{U}(r_{01}, m_1), \vec{U}(r_{02}, m_2))$ are shown for comparison. Force $\vec{F}$ on the target central atom (green ball) is indicated by the thick black arrow. The internal vectors are functions of the displacement vectors of the neighbouring atoms (red balls) while the parameters of $(r_0,\ m)$ are both adjustable to generate internal vectors accounting for different shells of neighbouring atoms.

same symmetry group with QM force and the atomic environment (ii) when $\vec{V} = 0$ for all pairs of (r_0, m), so is the target force due to the directional correlation. Force can well be predicted to be zero without the explicit performing the GP regression. This is especially meaningful when dealing with the highly symmetric configurations, as to be explored in Section 5.8.

5.3.3 Weight Function

The introducing of an appropriate weight function for each of the neighbouring atoms is the basis for the derivation of a covariance between the configurations to meet the requirement of high-precision prediction by Gaussian Processes. In Fig. 5.5, the weight functions $\omega_i = e^{-\left(\frac{r_i}{r_0}\right)^m}$ are illustrated by the radial cutoff curves where a pair of tuneable parameters (r_{cut}, m) are adopted to generate a set of correlated internal vectors. From the plot, it can be seen that r_0 corresponds to a critical point after which the weight function smoothly decreases from weight of 1 to 0 while the parameter m controls the steepness of the transition zone. Instead of using a sharp cutoff for the atomic

environments, we introduce a smooth weight function to geometrically screen the interactions from far-away neighbouring atoms. It is noted however, that r_0 is not the rigid cutoff for the local environments, but a transition point above which the neighbouring atom contributes much less significantly.

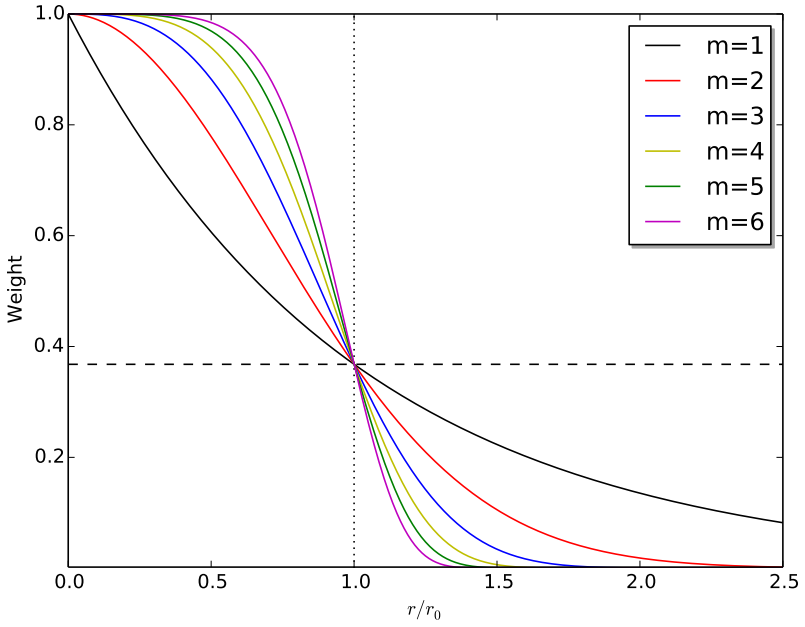


FIGURE 5.5: A plot showing the weight functions adopted to obtain internal vectors. Two tuneable variables (r_0, m) are adjustable while they are physically related to the cut-off set for the atomic environments and decaying power of the contribution from far-away neighbouring atoms. The vertical (dotted) and horizontal (dashed) lines mark the critical points $r = r_0$.

5.4 The Feature Matrix

So far, we have derived the set of internal vectors by symmetrically representing the features of the atomic environments. However, when numerically measuring the distance between two sets of vectors, the dependence of these vectors on the Cartesian reference frame is still problematic. To diminish the dependence while at the same time maintaining the feature-representing nature, we further construct the projection matrix from the internal vectors and use it to measure the difference between two atomic environments. We denote this projection matrix as 'feature matrix' in this thesis, and it is expressed

as $M \equiv V^T A$ where

$$V^T = \begin{pmatrix} | & | & | \\ \mathbf{V}_1 & \dots & \mathbf{V}_k \\ | & | & | \end{pmatrix}, \; A^T = \begin{pmatrix} | & | & | \\ \hat{\mathbf{V}}_1 & \dots & \hat{\mathbf{V}}_k \\ | & | & | \end{pmatrix}. \tag{5.6}$$

In the above equation, V^T contains the set of internal vectors while A^T gives the corresponding internal directions. The superscript T marks transpose operation on the matrix. Both vector sets are expressed by the Cartesian components. Feature matrix M derived in this way has the same symmetry as the atomic environments and the force vector. All the elements in M give an complete correlation between the internal vectors.

5.5 The Correlation between V_i and $\vec{F}_{QM}$

In this section, I present an analysis of the correlation between $\vec{V}_i$ and $\vec{F}_{QM}$. In statistics, correlation is a quantitated measurement of the dependence or concurrence of two random variables [96]. The internal vectors captures the property of QM forces in terms of both symmetry and locality, and interesting correlations exist between them. It is possible to incorporate using of any available, well-tested classical force vectors into the representation, which can further improve the efficiency and accuracy of the ML prediction, as to be discussed in detail in Chapter 6.

r_0 \ m	1	2	3	4
0.5	(0.079, -0.312, -0.049)	0	0	0
1.4	*	(0.396, -2.465, -0.355)	(-0.134, -0.385, -0.032)	(-0.008, -0.009, -0.001)
2.3	*	*	(5.00, -12.33, -2.258)	(3.631, -14.12, -2.176)
3.2	*	*	(8.709, -1.377, -1.072)	(10.82, -10.94, -2.537)
4.1	*	*	*	(3.631, -14.12, -2.176)

FIGURE 5.6: A table showing the scheme used to derive internal parameters based on the pair of parameters in the weight functions. The table shows the internal vectors corresponding to different pairs of (r_0, m), while the $*$ indicates those that have significant contributions from distant neighbouring atoms and are not suited for construction of the representation M. The vectors which have length smaller than a numerical threshold (10^{-6} in this table) are reset to be zeros.

I investigated the directional correlation between the vectors of $\{\vec{V}_i\}$ and the QM force $\vec{F}_{\mathrm{QM}}$ by using the correlation coefficient: $\mathrm{corr}(\vec{V}_i, \vec{F}) = \sum_{j=1}^{N_{data}} |\hat{V}_i^{(j)} \cdot \vec{F}_{\mathrm{QM}}^{(j)}| / \sum_j |\vec{F}_{QM}^{(j)}|$.

The correlation is calculated as the force projection onto the individual internal directions and target first-principles force vectors then averaged over an entire database, containing 4000 QM data configurations which were generated from canonical MD trajectory of Si at 1000 K, each marked with index j. The QM forces were calculated within DFT framework as implemented in the VASP package [25, 26] with the ultrasoft pseudo-potential approximation [24]. The calculated correlation factor is plotted in Fig.5.7 for a range of r_0 and m.

With quantitated correlations, procedures to optimise the representation can be performed and this favours the force machine learning accuracy. When choosing the representing internal vectors, two preliminary factors must be considered: (1) those correlating with the QM force are favourable (light-colored region in Fig.5.7). This helps to reduce the total representation dimensionality and accordingly increase the computational efficiency. (2) those independent from the influence of the far-away neighbouring atoms are favourable. Other consideration are regarding the completeness of the representation as well as the independence between two internal vectors.

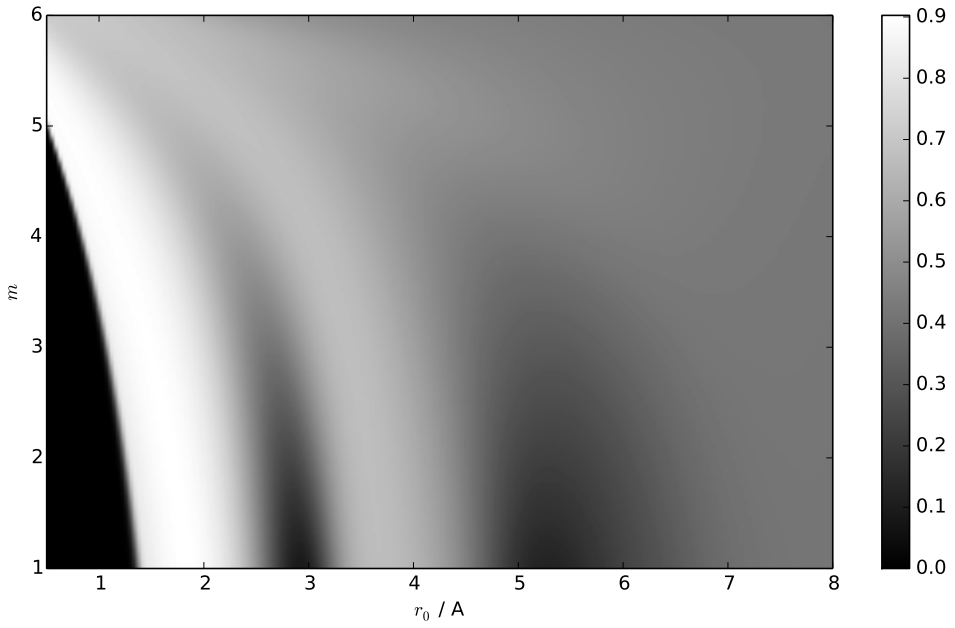


FIGURE 5.7: A plot shows the correlations between internal vectors and corresponding QM forces $\vec{F}_{QM}$, calculated upon a database containing 4000 configurations of Silicon, generated from the MD trajectory at 1000 K. The value of the correlation is scaled with respect to the $corr(\vec{F}_{TB}, \vec{F}_{QM})$.

5.6 Configuration Similarity

The distance between two atomic configurations (labelled as α and β) is evaluated by the distance of their feature matrix as expressed in Eq.5.7. In the equation V_i is the magnitude of the i-th internal vector with k being the number of internal vectors. X_{ij}^{α} indicates the projection of j-th vector $\vec{V}_i$ onto the i-th direction $\hat{V}_i$ for configuration α. The same applies for configuration β. Under this distance metric, two atomic configurations are considered to be identical (zero distance) only if they have strictly the same representation, in other words the same set of internal vectors.

$$d_{\alpha,\beta}^2 = \frac{1}{k}\sum_{i,j=1}^{k}\left[\frac{X_{ij}^{\alpha}}{\chi_i} - \frac{X_{ij}^{\beta}}{\chi_i}\right]^2 \tag{5.7}$$

The weighting factors $\{\chi_i\}(i = 1, \cdots, k)$ introduced in Eq.5.7 for each of the internal directions are meant to normalise the projections so that each of the internal vectors give equally weighted contribution to the distance measurement, as expressed by the following equation (N gives the number of configuration in the database):

$$\chi_i^2 = \sum_{\alpha,\beta}^{N}\sum_{j=1}^{k}\frac{(X_{ij}^{\alpha} - X_{ij}^{\beta})^2}{N^2}. \tag{5.8}$$

They can be derived by statistical procedures upon a given QM database, and their values depend on the domain of the configuration complexity. A complete representation of the geometry of neighbouring atoms is essential in this work, as any ambiguity in distinguishing two configurations could bring in large systematic error for the predictions. However, balance should be made between the completeness and cost of higher dimensionality. While less internal vectors cannot completely captures the difference between configurations, using more internal vectors risks to separate all the data points to be distant from each other. The fitted curves for them using an 'Kernel Density Estimation' (KDE) algorithm [97, 98] are seen in Fig.5.8.

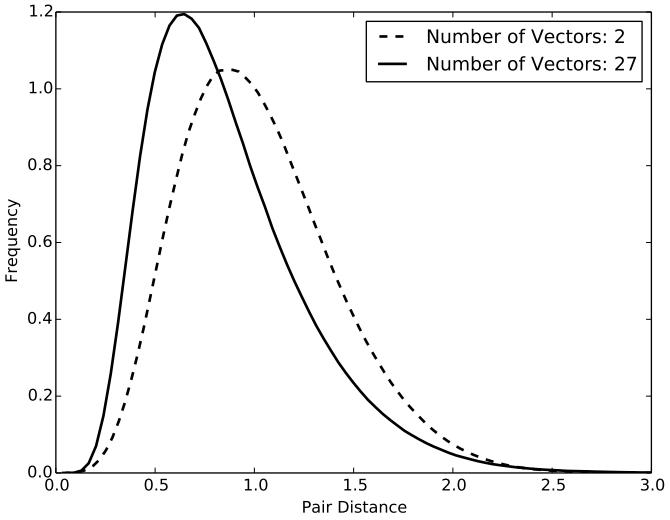


FIGURE 5.8: The pair distance are defined as the sum of each individual distance as in Eq.5.7. As the squared sum of each Gaussian distribution, the overall distance shows χ-squared-type distribution.

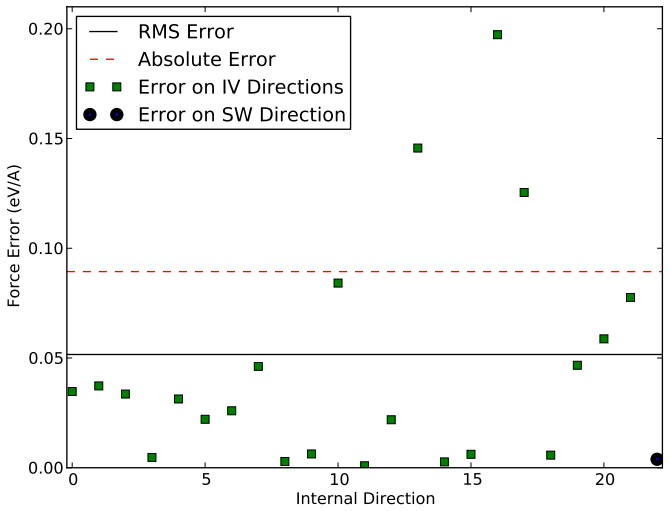
5.7 Over-determined Force Components

The k force components along each of the internal directions are predicted *via* the GP function inference with the same covariance matrix, however, based on the force component data on the corresponding internal directions. The predicted internal components do not make a single vector quantity in the external reference frame, but with Bayesian variance on each of the components. By the GP calculation, each component is of Gaussian distribution. From all the predicted components, the most-likely force vector can be computed by minimising the square residuals on all the internal projections $\min_X \|\mathbf{AX} - \mathbf{F}\|$, where $\mathbf{A}$ is the internal direction transformation matrix, $\mathbf{X}$ the force vector in Cartesian coordinate to be determined, and $\mathbf{F}$ contains the force components on all the internal directions. Under this least-square procedure, $\mathbf{X}$ can be expressed as,

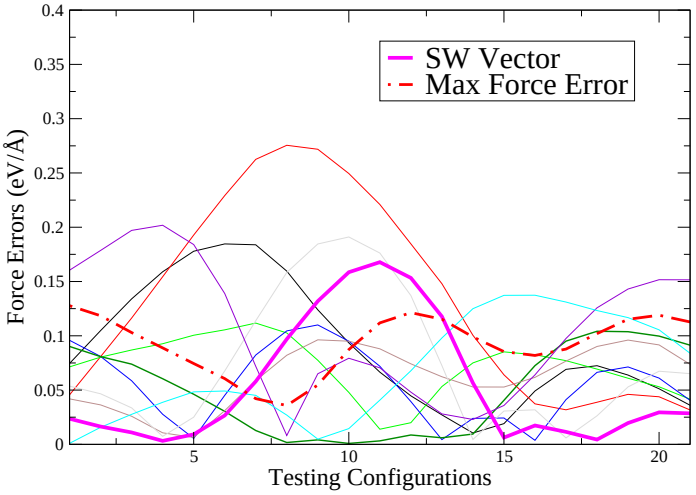
$$\mathbf{X} \simeq (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{F} \tag{5.9}$$

A simple illustration of the proposed procedure is given in Fig.5.9. The error made by GP prediction on each of the force components are different on each of the internal directions $\{\hat{V}_i\}$ where $i = 1, 2, \cdots,$ and k. The integrated force vector however, has better precision than the large individual errors. The significance is that the predicted

force vectors are not misled even though some predicted components may be far from the ideal value.



(a)



(b)

FIGURE 5.9: In panel (a): force error on each of $k = 23$ internal components. The 23-th components corresponds to the SW force direction, while the RMS error between the predicted force and the target QM force was marked by the horizontal line (see text for details). In panel (b): *test* along 20 fs MD trajectory, the force error from different internal directions and the Max force error derived from the least-squares solution of the over-determined equation. The error along the SW vector direction are also shown.

With the force components $\{F_i\}$ from ML predictions, the most-likely force vector $\vec{F}$ in the original Cartesian Coordinate can be derived from the mean value along with the

prediction variance which can be used as a proper weight bias on each of the internal directions. Alternatively, the certainty level on each of the internal directions can be assessed via a weight counting procedure. We have seen that the overall force error is much lower than the individual component error, which shows that the least-square approach can give good prediction precision and avoid the bad predictions on some individual directions. Knowing the correlation weight factor in each internal direction, in the further procedures, higher weight can be assigned to that direction while less on the others.

5.8 Highly Symmetric Configurations

Under the internal-vector representation, two associated problems have to be addressed, i.e. (1) degeneracy of the internal vectors (2) flipping of some internal directions.

The degeneracy of the internal vectors presents a hidden systematic error for the prediction. In this case, the procedure done using least-squares approach in Eq.5.9 does not work, as the direction transformation matrix A becomes lower-ranked and inverting of the geometry matrix $G = (A^T A)$ involves singular value (numerical overflowing) for the degenerated dimensions. Therefore the predicted force components are not adequate to restore the force vector in the 3D Cartesian space. In phonon calculation, for the distorted configurations from equilibrium crystal structure, this issue becomes prominent and thus significant to be solved.

We can consider the directional correlation between $\{\vec{V}_i\}$ and target QM force as discussed in Section.5.3. This provides a good foundation to infer that their distributions are expanded in the same dimensionality and the internal-direction degeneracy problem can thus be addressed with the numerical procedure as follows. A threshold is set to indicate the dimensionality reduction after the Principal Component Analysis (PCA) on the internal-vector set $\{V_i\}$, and the force component on the degenerated dimensions are accordingly set to be zero with no significant accuracy loss.

1. Find three orthogonal principal axis by PCA computation on the internal-vector set $\{\vec{V}_i\}$: explicitly calculate the covariance matrix of $(\mathbf{V} - \bar{\mathbf{V}})^T(\mathbf{V} - \bar{\mathbf{V}})$ where $\mathbf{V}$ is the matrix containing the internal vector set and $\bar{\mathbf{V}}$ indicates the corresponding

mean vectors. Three principal basis $\mathbf{X} = (X_1, X_2, X_3)^T$ are generated in the decreasing order of the eigenvalues $\lambda_i (i = 1, 2, 3)$.

2. In the PCA coordinates, dimensions with the eigenvalue λ_i smaller than threshold δ are artificially reduced while force components in those dimensions are set to be zeros. The non-trivial dimensions become solvable in either 2D plane or 1D line. Both internal vectors and forces are expressed in the new coordinates system of $\mathbf{X}$.

3. For the 1D case, forces are taken as average over the components along the single principal dimension. For the 2D case, the least squares solution is performed but on the two reduced dimensions, by which the force vectors in the Cartesian space can be extracted from the predicted 2D force components.

The other issue is the directional flipping of the internal vectors with trivial magnitude. Accompany with the flipping of internal direction(s), dramatic change can be seen in the projection components in the flipping direction(s) and thus the representation, while the variation of the actual atomic configuration undergoes insignificant changes. The GP prediction within this regime may have to involve large uncertainty due to the abrupt function relation. Numerical way to solve this problem by diminishing the contributions of the vector $\vec{V}_j$ to the feature matrix whenever its magnitude $\|\vec{V}_j\|$ becomes smaller than a threshold δ, as expressed below, where suitable value for the threshold can be obtained by testing on a given database.

$$\vec{V}_i \cdot \hat{V}_j = \begin{cases} 0 & \text{if } \|\vec{V}_j\| \leq \delta \\ \|\vec{V}_i\| \cos\varphi_{ij} & \text{if } \|\vec{V}_j\| > \delta \end{cases}$$

As been noted, the above numerical procedures are only valid provided good correlation between the internal vectors and the learning target $\vec{F}_{QM}$ in terms of both direction and magnitude. This in return adds constraints to the internal vectors we adopt in constructing the representation for the atomic environment.

5.9 Summary

In this chapter, an internal-vector representation for the atomic configurations was established incorporating the hidden symmetries of the machine learning target, *i.e.*, Hellmann-Feynman forces. These requirements from symmetry, completeness, high-dimensional data space and smooth correlations between the internal-vector representation and the QM force were addressed. The data topology is determined by the explicit representation scheme adopted and also relates to the accuracy that can be inferred from the GP predictions. In the following chapters, implementation of this force calculation scheme into the large-scale MD calculations will be explored.

Chapter 6

Results III: Machine Learning 'On The Fly'

6.1 Introduction

In this chapter, the ML force calculation scheme introduced in Chapter 5 will be tested on static configuration databases which are generated from the MD trajectory within the framework of DFTB (see Section.2.19) or accurately DFT, with the internal representation additionally incorporating the empirical or semi-empirical force vectors. In the later part of this Chapter, this force calculation scheme will be implemented into the MD simulation and the associated errors comparing to the first-principles benchmark will be analysed to reveal the predictive capability until a large size of QM database was formed.

6.2 Static Learning Accuracy

We first investigate the accuracy of our ML force calculation in the interpolation domain. To do this, a database is built using consecutive configurations spanned by a reasonable time interval along the MD trajectory while force prediction is performed upon the test configurations chosen from the middle point of two consecutive data configurations, as shown in Fig.6.1. With test configuration sampled in such a way, the measured force

errors provide an evaluation of the maximum error of the force calculation during the predictor-corrector cycle, in analogy to the 'LOTF' force errors shown in Fig.2.3.

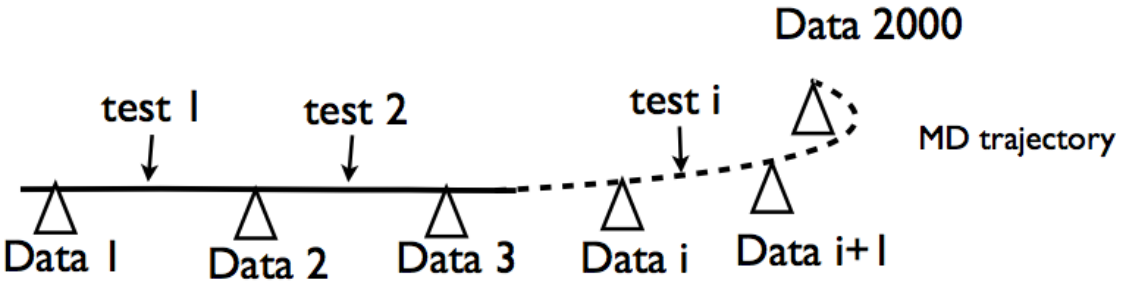


FIGURE 6.1: A schematic plot for generating the database from an MD trajectory run on the 64-atom bulk Si system. Data configurations in the database were collected at 20 fs intervals and test configurations from the middles of two consecutive data configurations.

For the calculations in Fig.6.2, 2000 data configurations from the MD trajectory [1] with time interval of 20 femtoseconds (fs) was used to make the database at the temperatures of 1000 K and 2500 K, respectively. The test configurations are from the middle points of two consecutive data points in the trajectory. In panel (a), the ML force accuracy is plotted for Silicon with respect to the teaching database size N_{teach} at the two temperatures. The teaching database are increased by the order of the similarity distance to the test configurations as calculated from Eq.5.7. Typically from the plot, the force error converges to 0.1 eV/Å, or relative error of 5 % at 1000 K and 0.25 eV/Å, or relative error of 10% for molten Si at 2500 K. It should be noted that, in the plot, the prediction errors at $N_{teach} = 0$ are equivalent to the average force magnitudes at two temperatures, as in this case of zero database, the predicted force for all configurations are constantly zero. We can see that, the average force magnitudes are $|\vec{F}|$ ~1.8 and 2.7 eV/Å for $T = 1000$ and 2500 K, respectively. In both temperature cases, the ML forces have much better accuracy than otherwise using classical Stillinger-Weber potential [43], which yields the error of 0.5 eV/Å or 30 % in relative error at 1000K and 0.9 eV/Å or 35 % in relative error at 2500 K, as marked by the dotted horizontal lines in the plot. The errors made by SW potential are literally close to those made by ML prediction only with several closest teaching configurations. Significant improvement upon this classical accuracy is thus achievable by adding more QM configurations into the teaching database. This

[1] The NVT MD simulation was run with 64-atom Si cell under PBC conditions with a 1 fs time step

data-based improvement makes the ML force calculation scheme differ from the function form fitting methods and is typically one of advantages holding by GP prediction.

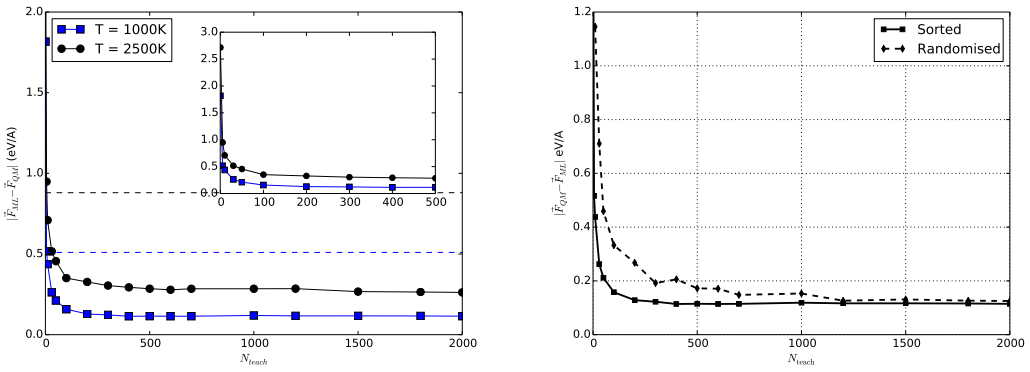


FIGURE 6.2: The ML-force accuracy test on DFTB force database. In panel (a) shows the error evolution for each test configurations with respect to the teaching database size, N_{teach}. The blue squares and black sphericals correspond to the temperatures of 1000K and 2500 K respectively, while the dotted horizontal lines indicate the average force error made by the SW classical potential at two temperatures (blue: 1000 K and black : 2500K). In panel (b) compares the different convergence rate of prediction accuracy with respect to increasing N_{teach} by sorted / randomised order of distance to the test configuration. Note that the classical SW force vectors were incorporated into the representation to achieve the best machine learning accuracy and efficiency (see Section 6.3 for force-vector augmentation).

The overall force accuracy from ML prediction systematically increases with respect to increasing the size of teaching database, for both hot bulk Silicon (1000 K) and melting Silicon at 2500 K. Snapshot of two data configurations at temperature from $T = 1000$ K and 2500 K are depicted in Fig.6.3. At 1000 K, the Si bulk system is distorted from the diamond structure while at 2500 K, there are significant changes in coordination type, enriching the database with much more three-fold coordinated atomic configurations. From our point of view of performing ML force calculation, this is a good example to show the learning capability for MD ensemble at different temperatures, which corresponds to the different sizes of accessible phase space.

In panel(b) of Fig.6.2, I give for comparison the increasing of teaching database size by completely randomised order in distance. The random ordering presents oscillated contribution from $N_{\rm teach}$. By sorting the teaching data according to distance with respect to the test configuration, more controllable accuracy was achieved and the accuracy typically plateaus at $N_{\rm teach} = 500$ teaching database. Considering the machine learning processes, there are two major factors that are related to the prediction accuracy: (1)

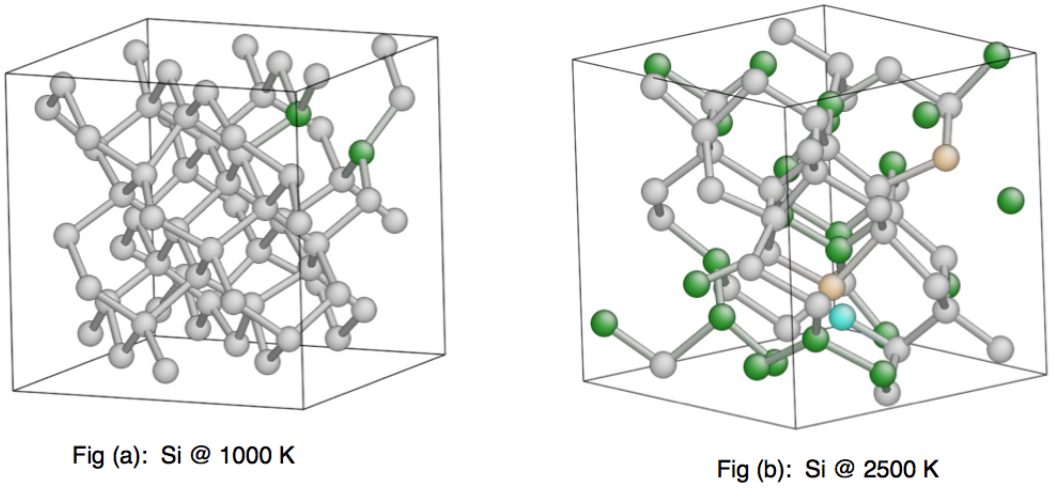


FIGURE 6.3: Snapshot from the database at 1000K (a) and 2500K (b) of Si for the ML force calculation. The different colors indicates the different types of coordination numbers. Grey indicates the four-fold coordination type, while green balls the three-fold coordination type.

the average distance of the test configuration to the data configurations (2) the size of training database N_{teach}. The convergence with respect to database size N_{teach} is found in the regime of smaller database, independent of the distance with respect to the test configuration. This is connected with the Bayesian nature of the prediction, i.e. *posterior* distribution has great dependence on the unveiled *prior* QM knowledge. However, optimal learning rate, $\mathrm{d}(\Delta F)/\mathrm{d}(N_{teach})$ can be achieved in the case of sorted data. Oscillation of the convergence takes place when the distance is in the randomised order, which signals the different weight of the contributions from close to far-away. This is derived from the correlation range between data points which is intrinsic for a given database. In our later ML force calculation, the sorted subdata are always used to replace the entire database to do the force teaching of the technique.

From Fig.6.2, we can see that the learning accuracy converges at $N_{\mathrm{teach}} \sim 500$. This reflects the local learning feature associated with the Gaussian kernel used to construct the covariance between configurations. It suggests that, for a given test configuration, a number of its closely relevant data configurations are sufficient to represent the entire database in the framework of GP prediction. In the MLOTF force prediction, especially when applied to the large-scale MD simulation, using sub-database sorted/selected (linear scaling factor) from a growing data repository are favourable in computational efficiency while having the best possible accuracy. These are very meaningful in the

dynamical ML and will be addressed in detail in the following chapters. In Fig.6.2, the accuracy converges at a typical value, which can be attributed to the following factors :

1. The correlation between the N-th teaching configuration and test configuration tends to zero as their distance increases. This is associated with the local learning feature, that the data configurations has less correlation with the test configurations becomes less weighted for the prediction accuracy.

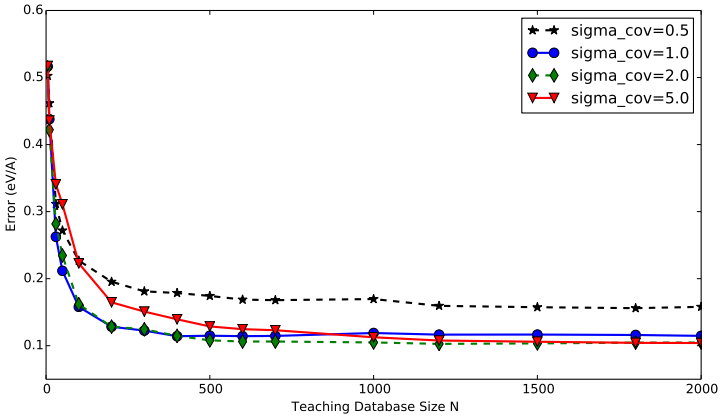
2. The representation itself has approximations. This may include the cutoff radius for the atomic cluster (8 Å in this test case for Silicon) and the completeness of internal vector representation. This error can be evaluated by a cross-validation upon the database. For a poor representation, the reproducing error (the error made on the teaching database itself) is much larger than the magnitude of the data noise assumed. However, this error prevents the further improvement on accuracy for the test configurations.

3. The error introduced on the QM database which is rooted in the DFT force accuracy itself, and was systematically smoothed out through a noise term: $\sigma_{\text{error}} = 0.05$ eV/Å that was used in these calculations, which is ultimate limit for the further improvement of the accuracy.

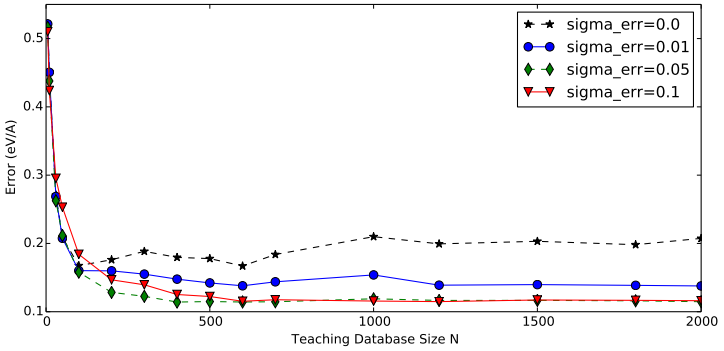
6.2.1 Hyper-parameters and Maximising Likelihood

There are two hyper-parameters we used to construct the covariance matrix for carrying out the Gaussian Processes, i.e. σ_{error} and σ_{cov}. Following the same interpretation of σ_{error} and σ_{cov} as in the standard Gaussian Processes [76], we illustrate the meaning of these two hyper-parameters in this context of ML force calculations. The topology of the database incorporating high-dimensional atomic environments is determined by the internal-vector representation as well as the hyper-parameters involved.

σ_{cov} gives the correlation length for the data pattern. As can be seen from Fig.6.4, larger σ_{cov} ($\sigma_{cov} = 5.0$) corresponds to longer correlation in the configuration distance and therefore more data points are required to obtain the converged learning accuracy. σ_{error} corresponds to an uncertainty exerted on each of the QM data and it is typically where Gaussian Processes differs from functional fitting. σ_{error} statistically is related to the variance associated with the predicted function from given database, as expressed



(a)



(b)

FIGURE 6.4: Plot (a): The accuracy of force prediction with respect to a chosen number of σ_{cov} for Silicon database at 1000 K. Plot (b) shows prediction accuracy with respect to a number of σ_{error}.

in Eq.4.10. Larger σ_{error} involves larger uncertainty but gives smoother prediction mean functional form. For smaller σ_{error}, the prediction is more similar to the over-fitting of data points, which is less useful for inference beyond the discrete data knowledge. The accuracy becomes diverged with respect to increasing the teaching database size, typically indicating the overused regulation from the data points, as seen from plot (b) in Fig.6.4.

Typically different from the empirical parameter fitting process, the hyper-parameters can also be numerically optimised by the approach of maximising the marginal likelihood. The hyper-parameters derived by Lagrangian parametric for the minimum problem as in (Eq.4.16) gives overall good performance in the calculations. For a given physical system, the optimal hyper-parameters are relatively localised and the optimisation can well be performed only when it is necessary. For large database, maximising the likelihood

can be costly. By selecting out the most closest subset of the database, dynamical optimisation of the hyper-parameters becomes feasible.

6.3 Acceleration for DFT Force calculations

In this section, I will give the prediction results upon a DFT database. Machine Learning of DFT or DFTB as the target learning object are equivalent in terms of function inference. However, they have underlying difference in the complexity of two different PESs, and smoothness of the two different force functions with respect to the represented configurations. The ML force accuracy is investigated further with augmentation of empirical force vectors into the representation.

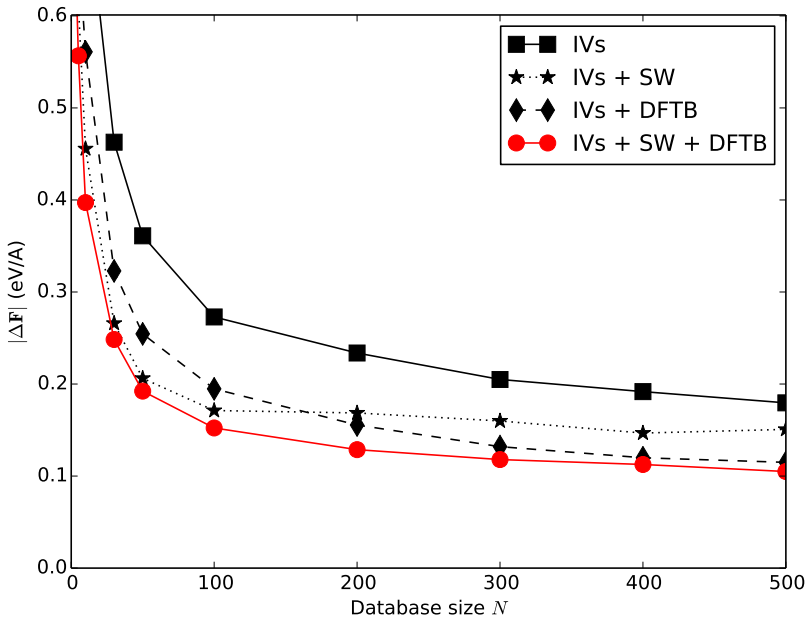


FIGURE 6.5: A plot showing the interpolation accuracy of MLOTF by testing the prediction error on configurations generated from the middle points of the data configurations during the interpolation cycles. A comparison is shown in the plot among different representation constructed with (1) Pure Internal Vectors (IVs) (2) Internal Vectors plus Stillinger-Weber Force vectors (IVs + SW) (3) Internal Vectors plus DFTB force vectors (IVs+DFTB) (4) Internal vectors plus DFTB force vectors and Stillinger-Weber force vectors (IVs+SW+DFTB).

The DFT database were generated using self-consistent plan-wave method, as implemented in the VASP package [25, 26]. In the ML process, the data configurations are sorted according to the distance with the test configuration. The predicted force error systematically decreases with respect to increasing the teaching database size,

N_{teach}. Similarly, ML force accuracy can be controlled around chemical accuracy of 0.1 eV/Å during the interpolation cycle, where data points are at time interval of 20 fs along the NVT trajectory of Silicon at 1000 K. In contrast, the average force error made by DFTB scheme is at 0.25 eV/Å while the average force error by SW is 0.5 eV/Å.

As a feature of our internal-vector representation, classical force vectors that have good matching with the target learning force can also be incorporated into the representation to improve the prediction accuracy. For ML of DFT database, this representation augmentation is studied incorporating non-DFT force vectors, for instance, the empirical SW force vector, or the semi-empirical DFTB force vector, or incorporating both them (SW + DFTB force vectors). The force accuracy for all of them are plotted in Fig.6.5 for comparison. The inclusion of $\vec{F}_{MM}$ vectors resulted in much faster convergence of the prediction accuracy $\mathrm{d}(\Delta F)/\mathrm{d}(N_{teach})$ with respect to the teaching database size, even in the case of augmentation with the less accurate force vector of $\vec{F}_{SW}$ which itself can be substantially deviates from the DFT benchmark. Further improvements are seen including in the representation more accurate vector: $\vec{F}_{TB}$ or including both SW and DFTB force vectors at the same time.

One further notable point from the plot is that, in the case of using $\vec{F}_{SW}$, accuracy is better at the small teaching database regime, while typically worse than using $\vec{F}_{TB}$ for the regime of large teaching database size (Fig.6.5 at $N_{teach} \sim 150$). As SW force fields for Silicon are only two-body accurate. For the small teaching database, or equivalently saying, for data configurations close to the test configurations, SW forces have very good correlation with DFT target. While the correlation is blunted as the data configurations from farther distance with the test configuration, usually three-body or higher terms dominate, randomness comes into disrupting the correlation between SW and DFT. We therefore see the accuracy get converged much faster than otherwise using DFTB which however, has better accuracy for longer range of interactions.

This systematic improvement of the prediction accuracy by incorporating additional force vectors into the representation is however not obvious, as the use of these apparently good vectors is no more than augmenting the existing set of internal vectors derived by Eq.5.5. In practical large-scale MD simulations at first-principles accuracy level, those well-validated classical or semi-classical force fields are desirably useful to enhance the

ML accuracy toward the ideal first-principles descriptions, with trivial increase on the computation cost than otherwise.

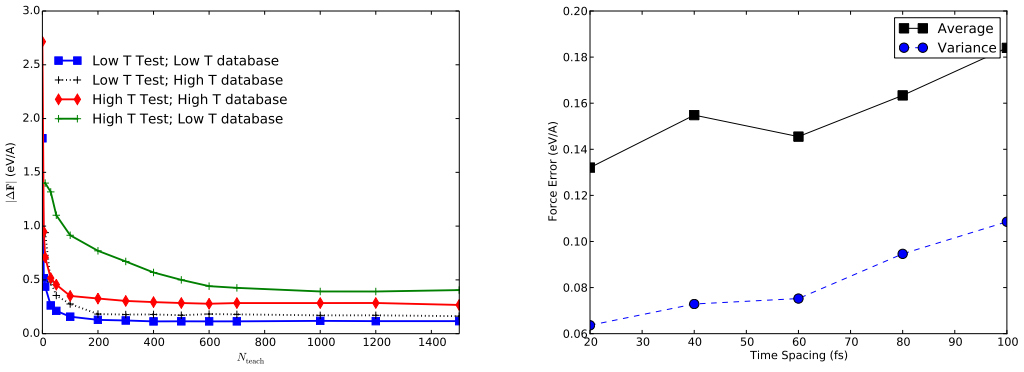
6.4 ML at Different Temperatures and Database Density

The transferability of any classical potentials or forces fields are a focused topic. Specifically for ML force calculation scheme, the transferability of the teaching database generated from different simulation projects or even under different chemical or mechanical conditions, is of importance in practical applications. To explore this point with this current force calculation scheme, the force accuracy was cross-validated with two independent databases collected from MD trajectory run at temperatures of 1000 K and 2500 K, respectively.

In Fig.6.6 plot the calculation results for comparison. For the test consideration, both the two independent databases are made up of 2000 atomic configuration. For the high-T (T = 2500 K) database, the data configurations are distributed over a larger area of the phase space than for the low-T (T = 1000 K) case. In Fig.6.6, high-T database evidently yields much better performance for force predictions on the low-T test configurations than the other way around (see the black dotted line in Fig.6.6). Significantly, this accuracy (0.12 eV/Å) has reached the level very close to that was achieved by prediction using the database from the same low-T trajectory. For the prediction on high-T configuration with low-T database, much larger errors are made than using high-T database, because of the inadequate QM knowledge in the database about the test configurations. This is consistent with the fact that GP function inference proves better performance in interpolation predictions than in the extrapolation for the configurations far beyond the knowledge of the existing database. This also explains the significant drop of force error in Figs.6.2 and 6.5 even with a few less correlated or extensive data configurations than the unmeaningful prediction from zero knowledge. A more systematic calculation for the machine learning and predicting along the MD processes with alternating temperatures can be found in Section 7.3.

Since the density of the data configurations in phase space is one key factor for the accuracy and efficiency of the force calculation, it was further investigated through calculation using a database sampled from different time spacings along the MD trajectory

while the size of teaching database is kept constant at $N_{teach} = 400$. The results in panel (b) of Fig.6.6 show that in the case of coarser sampling density, the overall force accuracy is not seen dramatically worse. The prediction accuracy is thus, not a single function of one or two closest data configurations (either in time scale or space scale), in which case, the prediction is more associated with the linear function inference. The contribution from the distant data configurations becomes lost in the case of sufficient close configurations, which is a general result consistent with all the above accuracy-testing plots. However, when less than enough close configuration are available, the knowledge inferred from the distant data becomes more significant, as suggested by the Fig.6.6. Moreover, in panel (b) of Fig.6.6, one anomaly point in the curve is seen for the data interval at 60 fs (although variance is much larger), where the average prediction accuracy notably better than that at 40 fs. This means that the close data configurations to the test configuration in distance can well emerge from the configuration far-away along the time scale in the trajectory. This from different perspective supports the idea of using dynamical database across the MD simulation scale, which induces the later parts of this Chapter 7.



(a) (b)

FIGURE 6.6: In panel (a), ML force accuracy was evaluated using databases from independent MD trajectories at two different temperatures, $T = 1000$ K (Low T) and $T = 2500$ K (High T). In panel (b), The ML force accuracy is explored for different teaching databases that are collected at different time intervals in the MD trajectory. The calculation was performed on the same test configurations for the reason of comparison. For each database ($N_{teach} = 400$), average force error (black squares) and their standard deviations (blue dots) are plotted.

6.5 Phonon Calculation

Phonons are of fundamental importance for studying material properties at quantum mechanical level (see Section 2.1.8). Applying the ML force calculation into the phonon calculations, we have to technically deal with numerical problems, as for all the representing configurations, internal vectors can be close to zeros in this case. Vector 'flipping' typically happens and brings into disruptive discontinuity by the sudden rotations of some small representing vectors. This problem was addressed and can be found in detail in Section 5.4.

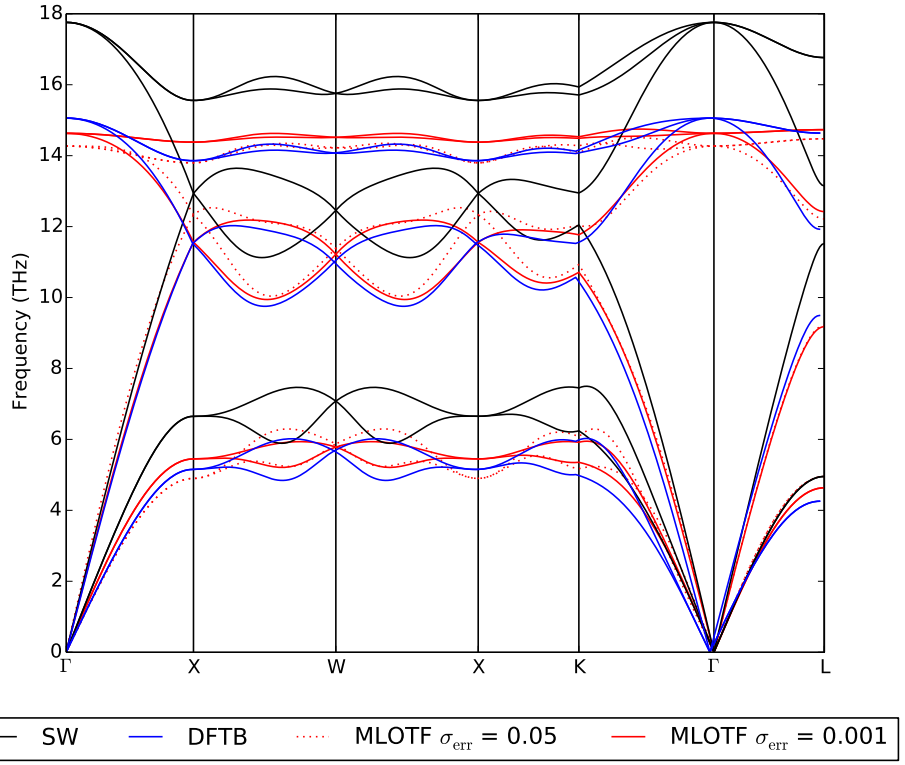


FIGURE 6.7: Comparison of the phonon spectra calculated by MLOTF (red, a=5.474 Å) and DFTB(blue, relaxed lattice constant a=5.474 Å), SW(black, a=5.44 Å). MLOTF phonon were calculated using $\sigma_{\mathrm{error}} = 5 \times 10^{-4}$ and 1×10^{-3} eV/Å. For the MLOTF calculation, database was 200 configuration generated from MD trajectory of Si at 300K. In phonon calculations, σ_{error} can be optimised by calculation at the Γ points.

The phonon dispersion curves are computed with the finite displacement supercell method using the PHONOPY package[99] to perform Parlinksi-Li-Kawazoe Fourier interpolation [30]. Phonons for the diamond-type Si at room temperature are depicted in Fig.6.7. Much better agreement were achieved with the DFTB benchmark than when using the

classical SW potential instead. The accuracy level with respect to varied hyperparameter of σ_{error} was also explored. It is worth noting that the hyper parameter σ_{error}, which controls the 'blurring' term used for regularisation of the prediction process. For high-precision force calculation like in phonon, smaller σ_{error} has to be used and at the same time, finer sampled database is usually needed for the required precision. Due to the limitation of using internal vectors, longer range interactions are in larger approximations thus harder to be described, which explains the larger discrepancy in the acoustic phonon modes than the optical phonon modes.

At the first-principles level of accuracy, phonon calculations are very expensive in terms of computer time, because self-consistent forces have to be computed for all the distorted configurations required by symmetries. This becomes particularly demanding for phonon calculations during MD simulation. Therefore, the application of MLOTF force into is worthwhile effort for accelerating the atomic force and accordingly phonon calculations. Also the incorporation of the technique into other alike research questions, such as time-dependent DFT, thermo-conductivity, heat diffusion etc is significant.

6.6 Computational Scaling

For Gaussian Process prediction with a dynamically growing database, a pronounced problem is the computational effort required to invert the large-dimensional covariance matrix C_N, which typically scales as $O(N^3)$, where N is the rank of C_N. To address this problem, one suggested approach is to use selected sub-database, keeping in mind the fact that the sorting/selecting algorithms have optimal scaling factor of $O(N_{\text{data}} \log N_{\text{data}})$, with N_{data} being the overall size of the teaching database. The database can be sorted in such way that only the most relevant configurations are selected for the GP teaching process. Since only N_{teach}-dimensional (typically around 500 in our calculation) matrices are involved in the inverting calculation, the force prediction becomes robust. There are three parts which are majorly involved in the time cost of the MLOTF force calculation: $T = T_{dist} + T_{sort} + T_{invt}$.

1. Calculating the pair distance: $T_{dist} \sim O(N_{\text{data}})$. This explicitly including two parts: the first is the pair distance of sub-database for constructing the covariance, which can be expensive as the number of pair distances to calculate is:

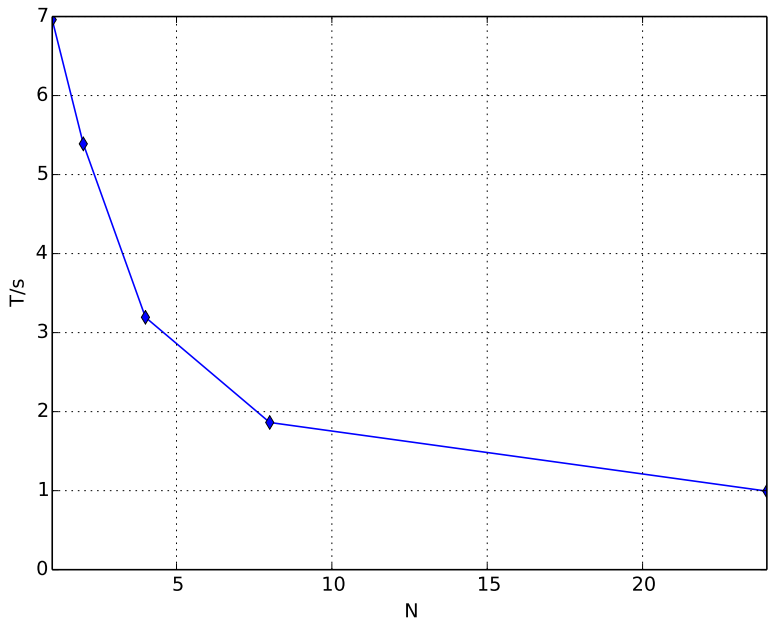


FIGURE 6.8: The efficiency of the Machine-leaning force calculation with respect to the number (N) of paralleled CPUs. Sub-database containing 500 configurations selected from the database were used for the Gaussian Processes. It is noted that, the calculation of T_{dist} dominants the times cost, as listed in comparison with other parts in Table 6.1.

$N_{teach}(N_{teach}-1)/2$. This can be in the order of magnitude of $\sim 10^5$ pairs for instance using $N_{teach} = 500$. This part of time cost however, can be diminished if we store the calculated pair-distance matrix into the computer memory, throughout the MLOTF force prediction process. The second part of pair-distance calculation is between test configuration and data configurations in the existing database. This part however has small pre-factor and linear scaling (see Fig.6.8).

2. Inverting the covariance matrix, $T_{invt} \sim O(N_{teach}^3)$ which can be very expensive for large teaching database, but for selected sub-database with typical size of $N_{teach} \sim 500$, the cost is insignificant.

3. Sorting the database and extracting N_{teach} most relevant configurations, $T_{sort} \sim O(N_{\text{data}} \log N_{\text{data}})$ with N_{data} being the size of the total database, and can be around the order of magnitude of 10^6. However, with optimal sorting algorithms, the process can be accomplished with no significant time cost compared to T_{discs} and T_{invt}, as listed in Table.6.1.

N_{CPU}	**1**	**2**	**4**	**8**
$T_{dist}(I)$/s	6.96	5.38	3.194	1.86
$T_{dist}(II)$/s	0.112	0.058	0.029	0.022
T_{invt}/s	0.086	0.086	0.086	0.086
T_{sort}/s	0.001	0.001	0.001	0.001

TABLE 6.1: The table gives the time cost for each parts of the MLOTF calculations as explained in the main text for comparison. T_{dist} are further divided into two parts: the pair-distance calculation between the data configurations ($T_{dist}(I)$) and that between the test configurations and each of the data configuration ($T_{dist}(II)$). In the calculations, only the dominant cost T_{dist} was parallelised and its scaling with respect to the number of processors was plotted in Fig.6.8.

Also worth noting is the algorithm for constructing the memorable relation for the database with the numerical strategy as implemented in the concept of K-D tree among the database with the computation time scaling as $O(N \log N)$ [100] (or sparcification of the database by constructing the hierarchy-clustering relation). The advantage is that, the relation can be stored and used generally for the existing database, and thus reduces the time in recalculating during each run of the force prediction. This is especially important for a database in the order of magnitude of millions. Based on the sub-database, optimisation of the hyper-parameters, for instance, 'maximising the marginal likelihood' can be performed without much computation time. The overall scaling for the MLOTF is $\sim O(N)$ and easy to be paralleled for calculations on large atomic systems.

6.7 Summary

In this chapter, the accuracy of the MLOTF on a static database from MD trajectory were tested and sorting/selecting the closest data configurations according to the similarity distance was adopted for the dynamical training for the large size of database. The incorporation of well-tested empirical force vector into the representation can enhance the ML prediction accuracy dramatically without significant cost of computational cost. In the above framework of prediction procedures, the overall scaling factor close to be linear and the calculation can be trivially paralleled. In the following chapter, application of MLOTF will be explored and the QM learning rate with rolling database across MD will be highlighted.

Chapter 7

Results IV: MLOTF Dynamic Learning

7.1 Introduction

The static machine learning accuracy has been explored in Chapter 6. During MD simulations, the incompleteness of the database and the renewability during the processes motivates the MLOTF in the MD simulations with a possible error indicator to update the database whenever necessary. In the following sections, systematic results concerning the MLOTF accuracy will be presented.

7.2 Application in MD Simulation

The flowchart in Fig.7.1 shows the scenario of the MLOTF molecular dynamics. During the simulation, efficient ML forces are used to replace the QM forces as long as the confidence level for the prediction is above a certain threshold δ, or predicted error smaller than threshold δ_{error}. For the configurations that are not predicted reliably (or predicted error $\geq \delta_{error}$), the QM routine will be called for to recalculate the forces and to augment the existing database. All the ML forces are thus predicted based on the best *prior* QM data available. Sub-databases are selected out from the sorting/selecting

procedure before performing the GP prediction. Under this scheme, dynamically updating the QM database is possible even up to the size of the order of magnitude of 10^6 configurations.

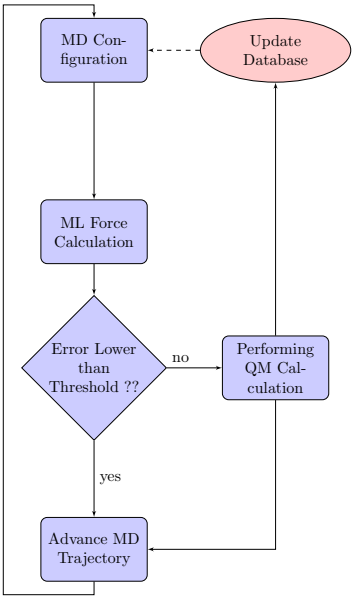


FIGURE 7.1: A flowchart showing the MLOTF MD calculations.

As an example to explore the MLOTF and the accuracy for the ML forces, MD simulations for a 64-atom bulk Si under PBC was set up. The QM learning target forces in these calculations are carried out by using the DFTB Hamiltonian for testing. The DFTB forces are calculated along with the ML calculation for each configuration to obtain a real error $|\vec{F}_{\mathrm{ML}} - \vec{F}_{\mathrm{QM}}|$, which is used as indicator for QM routines. In the later part of this chapter, I will introduce the possible approaches to derive the applicable prediction error, by using which, we can reduce the calling for the QM calculation to the point only when necessary. In Fig.7.2, MLOTF MD under different error thresholds were performed with database growing from scratch at $t = 0$ at the temperature of 1000 K. The QM calling rate $R(t)$ at each time step t was calculated by numerically averaging over the past trajectory from $t = 0$. This gives an evaluation of the overall efficiency that can be achieved by the MLOTF.

$$R(t) = 1/t \sum_{\tau=0}^{t} A(\tau) \tag{7.1}$$

where $A(\tau) = 1$ if a QM calculation is needed at time τ and $A(\tau) = 0$ otherwise. The QM calling rate $R(t)$ dramatically decreases at the initial stage of the MLOTF MD for

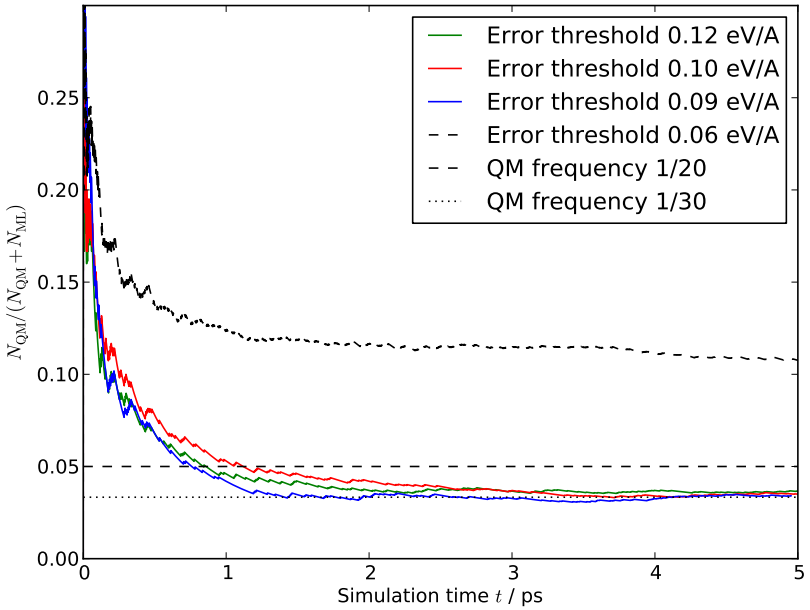


FIGURE 7.2: 'Machine Learning On The Fly' with real error monitored as the indicator to call the QM routines and feed the existing force database. DFTB Hamiltonian was used to perform the calculation of QM part as an efficient approximation to DFT. The converged QM calling rate signals the data coverage of the configuration space and highlights the existence of a typical core database representing the system.

all the thresholds. The smaller threshold produces higher precision ML forces along the MD and trajectories closer to the learning target. However in the case of using the error threshold of $\delta = 0.06$ eV/Å, the QM calling rate $R(t)$ falls to much higher value than using other thresholds $\delta_{error} \geq 0.09$. This is because the threshold 0.06 eV/Å has a value close to the data noise: $\sigma_{\text{error}} = 0.05$ eV/Å assumed in the GP inference and the prediction accuracy is informatively unattainable by simply adding more relevant data into the database. To achieve higher precision, a smaller σ_{error} has to be used and at the same time, more close teaching configurations are required to make the ML prediction, which however, limits the extrapolation capability of the overall MLOTF process. For a threshold δ_{error} around 0.1 eV/Å, as can be seen, averaged extrapolation time of 30 fs was obtained in the MD calculations up to 5 ps. This is a significant improvement with respect to the predictor calculation in previous non-learning LOTF. In these MLOTF calculations, we start from database each time from scratch, while in practice, the database can be cross-used from different simulation runs, thus the accuracy/efficiency can be further enhanced.

A plot for the ML efficiency at two different temperatures are given in Fig.7.3, where the

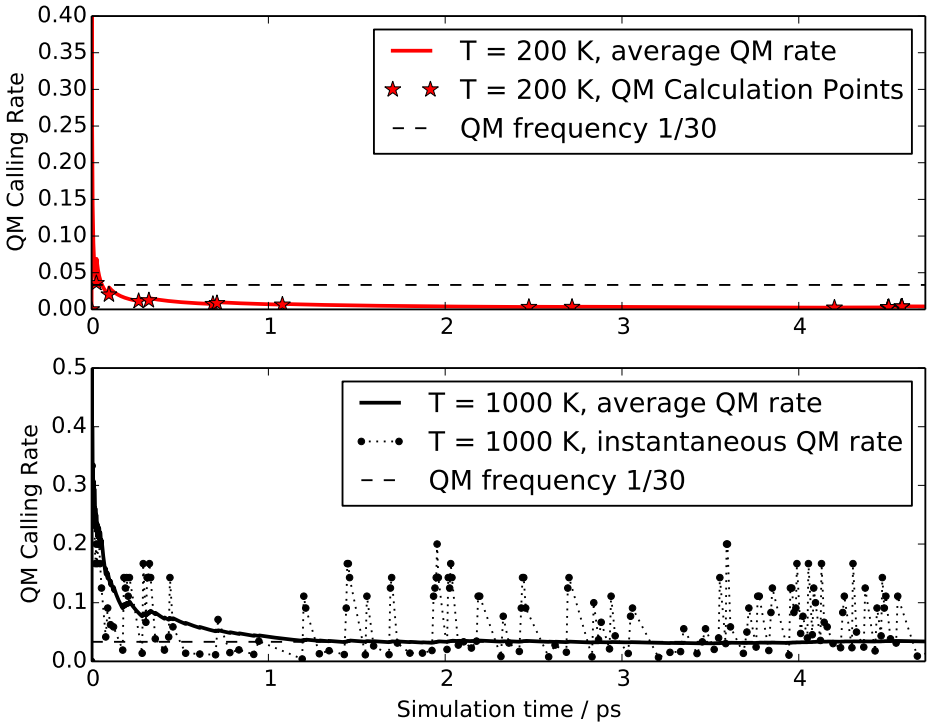


FIGURE 7.3: The average and instantaneous QM calling rates during the ML-force driven molecular dynamics. The QM calculation sites are marked by the red dots for T=200 K (upper panel). Both average and instantaneous rate are plotted for T=1000 K (lower panel). The error threshold is 0.1 eV/Å in the MLOTF calculations.

individual QM training sites in the MD trajectory are also marked by the red stars (upper panel in Fig.7.3). The instantaneous QM rates (black dots) are calculated as the inverse extrapolation time between each training site and the last training site in the trajectory. It gives an evaluation of the ML capability at the local part of the trajectory. At each of machine learning sites, all 64 atomic configurations in the periodic crystal cell are calculated quantum mechanically and added into the database. The average rate of QM calls fall to $\sim$ zero for $T = 200$ K after the initial pico-second (ps). The instantaneous rate of QM calls show that ML force calculation sustaining up to 2.6 ps with no need to perform new QM calculation can be seen in the MLOTF calculation, e.g. from 2.6 - 4.2 ps in the plot. For higher temperature $T = 1000$ K, the structural complexity becomes more significant and the QM calculations are more frequently required. The average ML extrapolation steps better than 30 fs were obtained, even though the instantaneous QM rates suggests that the structural complexity can be significantly and continually increases beyond the knowledge of the existing database till 4ps.

With a reliable error indicator in MLOTF, expensive QM calculations can be performed no more frequently than that required by a certain threshold in the practical MD simulations. The force calculations along the MD can be mainly carried out by the efficient ML process while the database is trained by QM routine. Under this scheme, the force accuracy in the MD is bound to the target QM level. Also interestingly, a useful representative database for a specific trajectory is resulted by the MLOTF procedure, which can be transferrablely used to simulations of relevant yet more complex chemical situations. Noteworthily and significantly, quantifying of degree of chemical novelty / complexity is provided by the rate of the QM calls revealed from the MLOTF simulations.

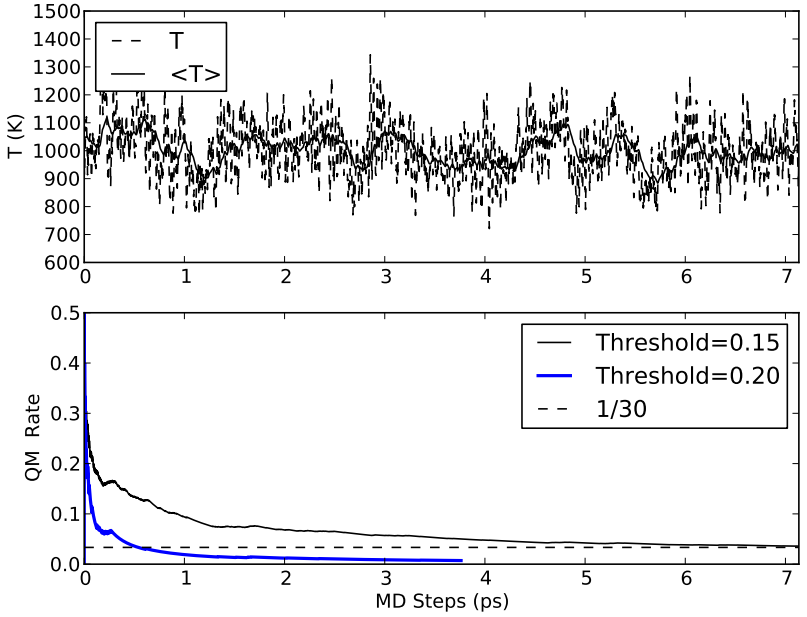


FIGURE 7.4: The upper panel indicates the temperature associated with the MD trajectory, including both the instantaneous temperature T and average temperature $< T >$ around 1000 K. The lower panel shows the ML force error for predicting on the given DFTB trajectory, with the real error used as an indicator for where to update the database. The dotted horizontal line gives an idea of the level of continuously 30 fs ML force calculation with no need to call for QM calculation.

Extended to longer time scale, MLOTF calculations were performed upon an NVT MD trajectories which were generated under the DFTB Hamiltonian at the temperature of T = 1000 K. The ML force accuracy was computed against the target forces for each step and the force errors were used as indicator for the QM calling. The results are plotted in Fig.7.4 where two different error thresholds are adopted: 0.15 and 0.2 eV/Å. In the plot for error threshold of 0.15 eV/Å, the average QM rate: $R(t)$ systematically decreases during the 7 ps trajectory, which indicates that the machine-learning capability

is dynamically improvable with more training data, and which is sign for the promising applicability of the methodology till very large database scale. As mentioned in Section 6.2, the limits from different channels (e.g. limits in representation, σ_{error}, the correlation, etc) become prominent for large-scale database, thus, optimisation procedure will be necessary to achieve better performances in this situation. Under even higher error thresholds (0.2 eV/Å), a rapid decreasing of the QM calling rate can be found during the initial 1.5 ps and QM calculations are thus not needed till really novel structural complexities are encountered in the future part of the simulation ($>$ 7 ps).

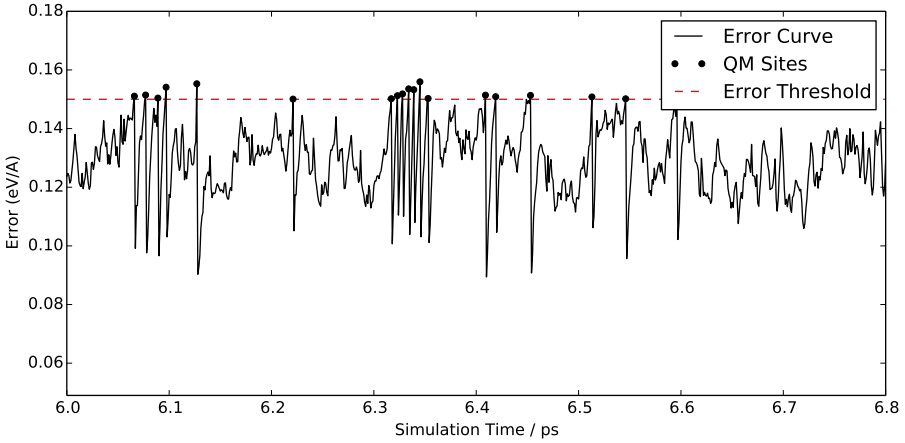


FIGURE 7.5: Snapshot from the prediction along the DFTB trajectory of Silicon at 1000 K and the extrapolation forces were predicted with sub-data set and error threshold of 0.15 eV/Å (indicated by the red dotted line), at a teaching database size N_{teach}=500. The typical zigzag (or sharp dropping down) shape in the error curve is a feature of this dynamic machine-learning scheme. Time scale (x axis) for this plot corresponds to a snapshot of that in Fig.7.4.

Due to the dynamical machine learning feature, the predicted forces are not smoothly evolving as from one single PES, so does the force error with respect to the QM benchmark. For T= 1000 K, a snapshot of the error evolution of the ML calculation along the MD is given in Fig.7.5. As seen, during the extrapolation region using only ML force, the average force error evolves with an oscillating fashion. For each force error above the threshold, QM training was performed to update the database. With the refreshed database, the ML force error restores to the minimum, which are around 0.1 eV/Å and typically larger than the σ_{error} used in GP prediction. This is due to the variation of the atomic environments under kinetic distribution along T = 1000 K along the dynamics trajectory.

7.3 MLOTF at Alternating Temperatures

The transferability of the QM teaching database to different atomic conditions is vital in large-scale simulations. In the static test given in Section 6.4, we have seen the prediction capability using the MLOTF at different temperatures. In this section, I describe a model using MLOTF upon trajectory segments generated from NVT MD under two alternating temperatures. The machine learning of QM forces during this process is depicted in Fig.7.6, along with the plotted alternating temperatures between low-T ($T = 300$ K) and high-T ($T = 800$ K) regimes. Note that the MLOTF calculation was performed with a single database both for high and low-T cases.

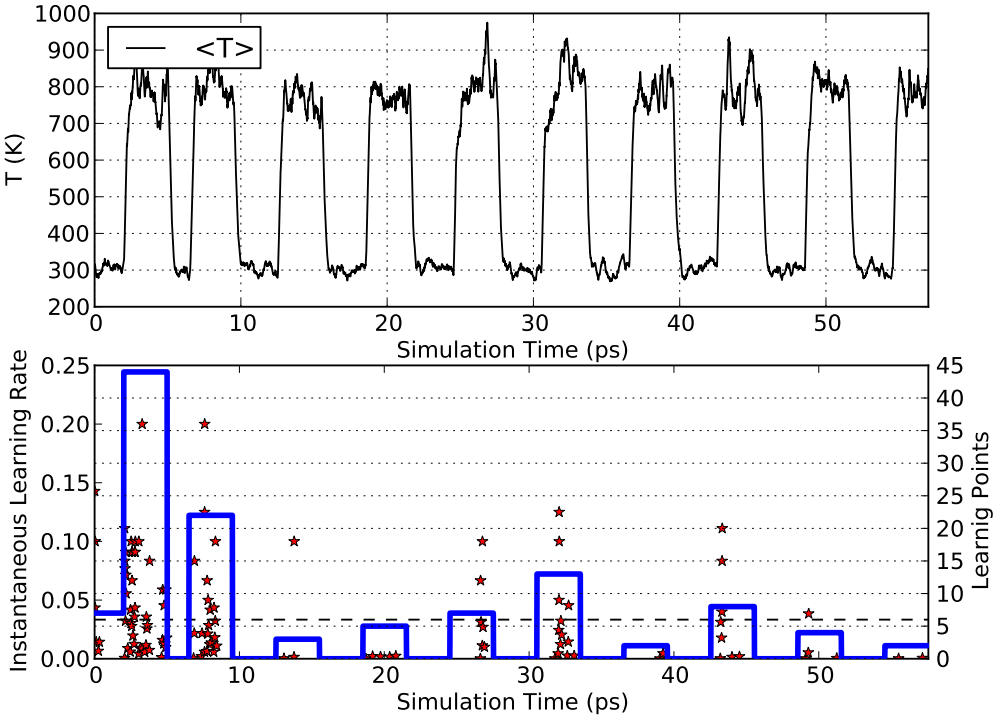


FIGURE 7.6: MLOTF calculation on the Si system to test the 'memory' and transferability of the learning informations across the different temperatures. In the upper panel gives the simulation temperatures and the lower panel illustrates the QM sites and both average and instantaneous QM calculation rates. The stars indicate the instantaneous learning rate, and the solid blue lines marks the number of learning points for each of the temperature segments. The error threshold for the MLOTF was chosen at 0.15 eV/Å for both the high and low temperature segments.

Each of the low and high temperature segments are 3 ps in duration after the first round of MLOTF calculation. We can see that intensive QM training takes place during the first round both for low- and high-T, while the database grows from scratch. Consistent

with the MLOTF for the static trajectory (Fig.7.4), the QM database gets saturated for the low temperature case ($T = 300$ K) rapidly after the first round of data training (around 50 trajectory frames, each containing 64 or equivalently in total 320 atomic environments). In the figure, the blue lines in the lower panel indicate the total number of QM training points needed for each of the temperature segments. For the second round of calculation, both for high and low-T cases, the number of QM learning points drops down rapidly, as a large amount of data configurations were ever learned from the previous round.

The database becomes complete for the ML force prediction at 300 K after the first round of QM teaching and thus no further QM points were ever found in the later MLOTF calculations. Interestingly for the high-T case, though the overall learning rate decreases along the MD trajectory, intensive QM training reoccurs at the points where the temperature presents so strong oscillation that it can go much higher than the target thermosetting temperature of $T = 800$ K, for instance at the simulation time of $\sim$ 27 ps and $\sim$32 ps in Fig.7.7. This temperature fluctuations are connected with the emergence of novel atomic configurations that are not so far predictable using the existing database. These novel configurations have dramatic difference in the bonding variation and/or coordination type from the data configurations. This is physically accompanied by the local melting in the system under the strong thermostat used in this simulation, for which we used Langevin thermostat with the damping parameter of $\gamma = 0.02$ fs^{-1}. A recalculation for the piece of trajectory between 24.5 and 27.5 ps, yet with milder thermostat ($\gamma = 0.01$ fs^{-1}) evidently smoothes out the temperature oscillations (Fig.7.7) and accordingly requires much less training points (only one QM learning point was found at 27.256 ps in the calculation) throughout the high-T segments. From Fig.7.7, we can also see the revisiting of new local minima of energy landscape which was activated by the high-T kinetics.

For the segments afterwards, as more knowledge has been gained about the phase space at this temperature, few QM calculations are required along the MLOTF calculations and the learning rate shows a systematic decreasing until 60 ps, where the trajectory ends. Another perspective to look at the distribution of the temperature that QM learning takes place is given in Fig.7.8. We can see that most of the learning points are located at the high-T region and especially new configurations come up during the temperature uplifting from low-T to high-T zone. During the cooling process, since

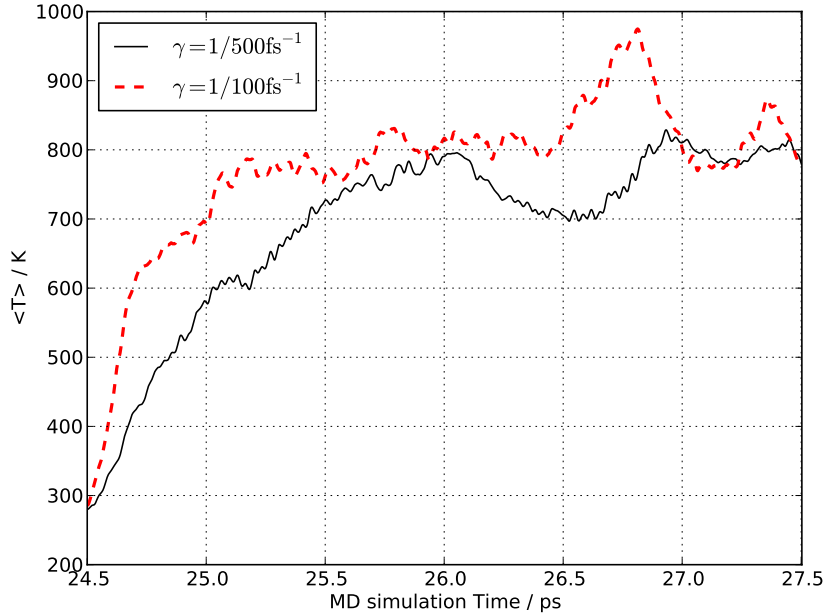


FIGURE 7.7: Two thermostats with different strength were adopted for comparison of the trajectory. The stronger thermostat $\gamma = 1/100$ fs^{-1} (red dotted line) case corresponds to the region from 24.5ps to 27.5ps in Fig.7.6. Using a milder thermostat γ=1/ 500 fs^{-1} (solid line), the temperature variation becomes smoothed out. Both trajectories were calculated starting from the same initial configurations.

structures are confined to smaller area close to the equilibrium, less learning points are thus needed during this processes.

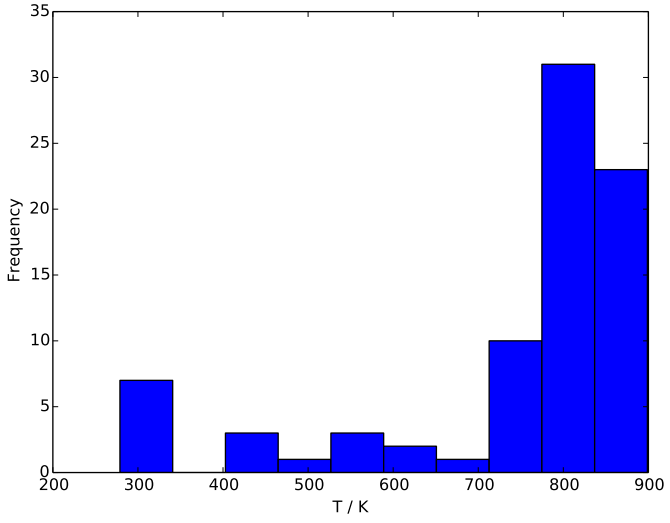


FIGURE 7.8: Distribution of the teaching points along the transient temperatures during the switching between two temperatures (300 and 800 K).

The machine learning and predicting was all from scratch in the above tests. In practice,

the database generated at different temperatures can all be made use of when probing into new chemical situations. The ML forces can be implemented in a predictor-corrector manner, where a possible future configuration along the trajectory will be calculated by QM routine to update the database, and a corrector for the former predictor cycle are performed with the renewed database. By this means, the overall efficiency can be improved by an extra factor upon the previous results without the corrector while the actual force errors can be restricted by incorporating a possible future configuration along the trajectory into the database *prior to* the ML prediction.

In addition to the predictor-corrector implementation, there is the other scheme for the MLOTF, with a prediction error navigating the new data feeding and QM calculations are only performed whenever necessary. The key issue in this application is the derivation of a faithful way to evaluate the prediction error, which can be from the variance term coming from the statistical product of Gaussian processes (Eq. 4.10). In the following section, further discussion of the prediction error will be presented.

7.4 Real Extrapolation with QM Database

In practical MD simulations, doing force extrapolation with MLOTF is appealing provided a good error indicator. One requirement for this error indicator is that it has a good correlation with the real error, i.e, $\Delta F = |\vec{F}_{ML} - \vec{F}_{QM}|$. The first possible way to estimate the prediction error is using the variance σ_{gp} that is obtained from the Gaussian Processes of Eq.4.10. As explained in Section 4.1.3, the Bayesian variance provides an indication of the uncertainty associated with the predictive mean value. Machine learning with coarse error threshold finds that the overall accuracy saturates rapidly along the MD trajectory. However, to reach higher precision, the coordinating of different data across large areas of the configurational phase space have to be coped with subtly.

In the MLOTF force calculation, there are k projection components are predicted from the GP process and each of them has variance of $\sigma^2_{\text{gp}} = \kappa - K^T C_N^{-1} K$, so that the forces are determined according to, $\mathcal{F}_i \sim N(F_i, \sigma^2_{\text{gp}})$ where F_i is the mean prediction along the channel i $(i = 1, \cdots, k)$.

Since the reconstructed force vector in the Cartesian space are written as:

$$\vec{F}_{3\times 1} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathcal{F}_{\mathrm{k}\times 1} = G_{3\times \mathrm{k}} \cdot \mathcal{F}_{\mathrm{k}\times 1} \tag{7.2}$$

with $G_{3\times \mathrm{k}}$ corresponding to the geometry factor which is associated with the transformation to reconstruct the force vectors in Cartesian Coordinates from the set of over-determined projection components. Based on the Gaussian Processes, the *posterior* distribution is also Gaussian distribution. By the rule of error propagation with the linear transformation in Eq.7.2, the yielded variance for each of the three Cartesian force components are:

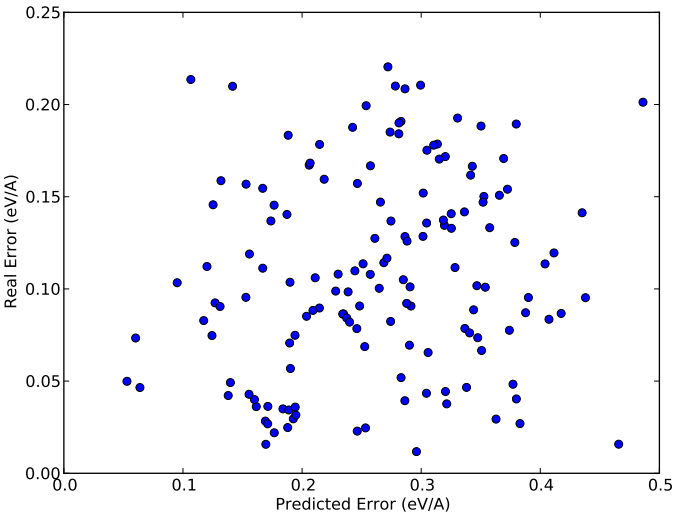


FIGURE 7.9: Error correlation between the prediction error and the real error with the prediction error derived from the variance of the Gaussian Processes, i.e. σ_{gp}

$$\sigma_{f_i}^2 \sim \sigma_{\mathrm{gp}}^2 \sum_{j=1}^{k} |G(i,j)|^2 \tag{7.3}$$

where $i = 1$, 2, and 3. The correlation between predicted error defined in the above way and the real error are shown in Fig.7.9. The correlation with real error is not satisfactory. This is rooted in the high-dimensionality of the configuration data type which makes the error harder to predict than the one-dimensional case and each of the projection components are not strictly uncorrelated Gaussian. Nevertheless, a typical pattern associated with the Gaussian was confirmed: for the case of small prediction error, the actual error is much less likely to be distributed within the very wild error zone (or say, real error larger than 0.2 eV/Å). The prediction error can overestimate the error more likely, while the other way can also be true, though less likely

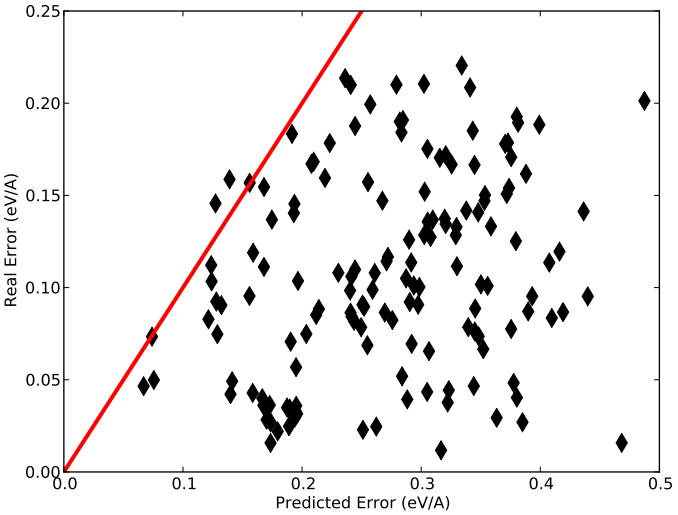


FIGURE 7.10: The correlation between the prediction error σ_{max} and the real force error. This can be useful in accurate atomistic simulations based on the upper bound of the indicator for the expensive QM calculations. The red solid line marks the boundary where prediction errors are equal to the real error.

To do force calculation with best possible prediction error in this methodology, we can adopt the maximised variance for the predicted force derived from the Gaussian variance σ_{gp}.

$$\sigma_{\max} = \sigma_{\mathrm{gp}} \cdot (\sum_{i,j=1}^{k} |G(i,j)|) \tag{7.4}$$

The correlation between this predicted error $\sigma_{\max}$ and the real force error: $|\vec{F}_{ML} - \vec{F}_{QM}|$ is shown in Fig.7.10. Blank area typically appears above the boundary read line that marks the ideal correlation between predicted error and the real error. This plot suggests that with the error indicator in Eq.7.4, the MLOTF force calculation can be strictly confined to be close to the QM benchmark.

7.5 Summary

In this chapter, the developed MLOTF force calculation scheme was justified in the large-scale MD simulations. The learning accuracy and efficiency along the MD trajectory was explored in Si system at different temperatures. The learning accuracy is systematically enhanced with the database growing as well as MD trajectory visit larger areas of the phase space. Given an accuracy threshold, the learning rate decreases as the database

grows, and new QM data learning is only needed when novel chemical environments are encountered in the simulation run.

Though for the current calculations, the representation for atomic environments is not completely optimised, the convergence of accuracy with respect to the number of teaching data configurations is confirmed. For some configurations, the maximum force error of MLOTF can be much bigger than the average value, however, with a prediction error indicator (either from the GP variance or from maximised prediction error), these configurations are to be calculated with QM routine to augment the existing database.

Though an ideal prediction error is not yet available, estimation of the upper bound of the uncertainty is possible and practical by considering all channels of error from the Bayesian inference, which are very useful for guiding the prediction process towards DFT accuracy. Also with predictor-corrector algorithm, MLOTF can be performed by increased interval along the MD trajectory and also the occurrence of large force error can be restricted in this way.

Chapter 8

Preliminary Results on Binary System

In the following, I will present the preliminary work carried out to extend the MLOTF force calculations into more complex material properties, such as binary system of SiC and SiO_2.

1. SiC is prototype system in the binary crystal family and it is widely-used in industry due to its excellent mechanical properties. There are two structure types of SiC at ambient conditions: $\alpha - \text{SiC}$ is hexagonal structure type and $\beta - \text{SiC}$ has structure similar to the diamond structure type [101, 102]. We shall consider the $\beta - \text{SiC}$ in the following. For SiC, the representation has to account for two distinct species and encode the complete information about the structure. The internal vectors $\{V_i\}$ should be so constructed that have optimal correlation with the target force vectors. To address the problem, we derive two separate groups for each of the species, i.e., S_1 for Si and S_2 for C in the case of SiC, while the feature matrix is comprised of different blocks, with the two diagonal blocks lists the contributions from the two different species and the off-diagonal blocks represents the joint contribution of the two species, as shown in the following:

$$M = \begin{pmatrix} S_1 \cdot \hat{S}_1 & S_1 \cdot \hat{S}_2 \\ S_2 \cdot \hat{S}_1 & S_2 \cdot \hat{S}_2 \end{pmatrix}$$

where $S_i \cdot \hat{S}_j$ marks the projections of the vector set of i-th species onto the vector direction set of j-th species. The internal vectors for species-i are the sum of the bonding

directions from only positions of species-i with appropriate weight function depending on the inter-atomic distance. The i-th internal vector in the vector set S_j is evaluated by Eq.8.1, and the $\delta(s_i - s_j)$ means that only when the neighbouring atoms of species s_j are taken into account in the sum. $\omega(r_i)$ the same weight function taken as before in Eq.5.5, and the sum takes place for neighbouring atoms within a cutoff radius.

$$\vec{V}_{i,s_j} = \sum_{i=1}^{Neighb.} \hat{\mathbf{r}}_i \cdot \omega(r_i)\delta(s_i - s_j) \tag{8.1}$$

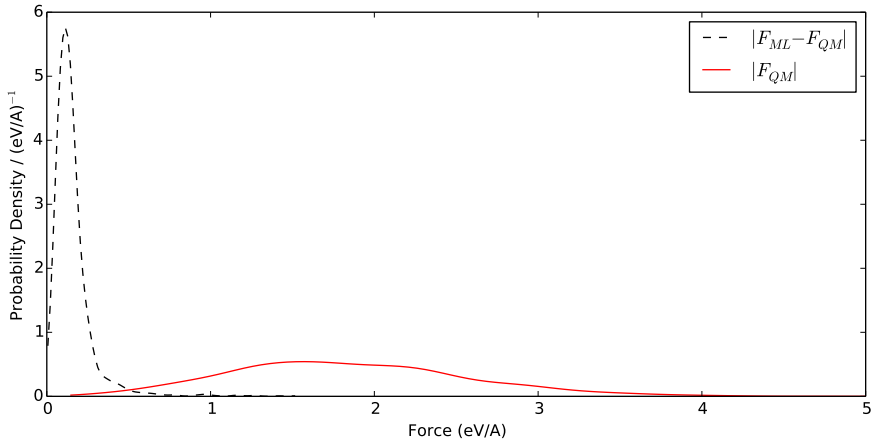


FIGURE 8.1: Plot shows the MLOTF force accuracy for SiC. The predictions were performed on 2000 configuration from trajectory that was performed for a diamond-structure SiC under the DFTB Hamiltonian. The error distribution was fitted by using the Kernel Density Estimation (KDE) [97, 98].

The results for SiC based on a database generated from an NVT trajectory at 300 K using the DFTB Hamiltonian. As given in Fig.8.1, comparing with the QM force magnitude, relative force error are less than 10 %. Even with the representation of 12 internal vectors, the learning convergence for SiC at 1000 K shows significant improvement on the learning capability when more data configurations are incorporated into the database, as seen in Fig.8.2. In this case, much longer range correlation between test configuration and data configurations was found with respect to increasing the database size up to $N_{teach} = 2000$, which is associated with the increased structural complexity and also suggests the incompleteness of using these 12 internal vectors as a representation.

2. SiO_2 For SiO_2, the long-range ionic bonding nature means much more internal vectors are needed to completely describe the local environments and high dimensionality

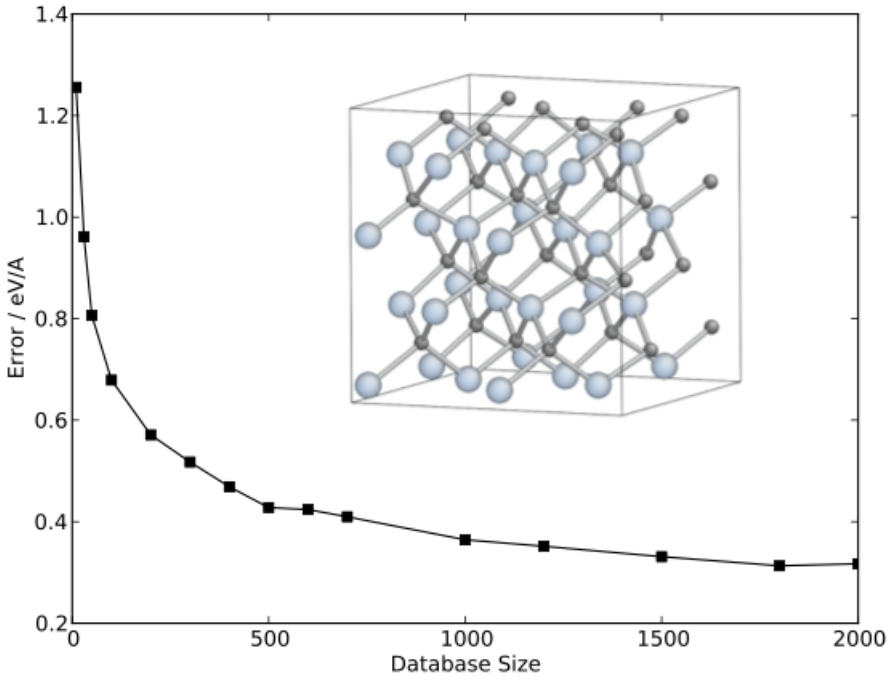


FIGURE 8.2: The force error convergence with respect to increasing the database size for SiC at 1000 K. As a preliminary calculation, the adopted representation includes 12 internal vectors with atomic cluster cutoff radius of 6 Å. The insert shows the structure of SiC with grey and dark balls indicating the Si and C atoms, respectively. Charge transfer has to be taken into account in SiC when doing ML force calculation from carved cluster, which makes it more challenging to represent the local atomic environments. Note that, no classical force vector was used in the representation for this calculation.

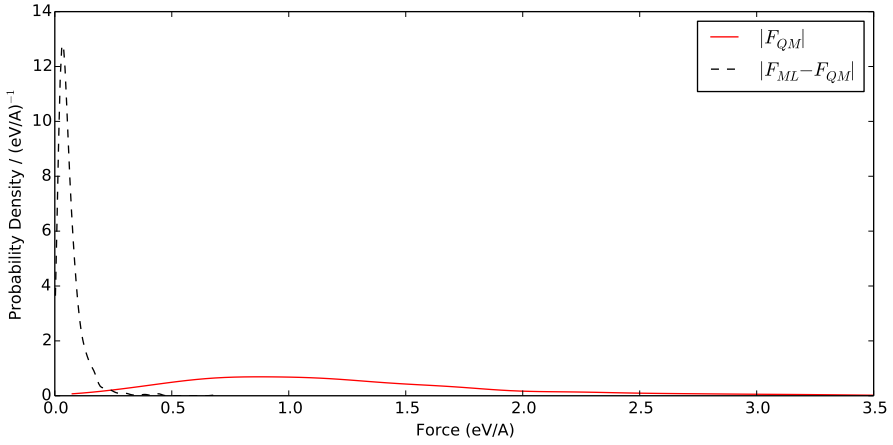


FIGURE 8.3: Error distribution of the ML force against the target QM force magnitude in Silica and the curves are fitted using the KDE algorithm.

of the data space. When charge transfer becomes prominent properties, the ML could be carried out upon the force after subtracting the best guess (such as the Tangney-Scandolo

(TS) interatomic force field [8]) for the long-range electrostatics. Here, I present some preliminary results of MLOTF calculation on this system. With atomic environments cutoff at radius of $r_{cut} = 8$ Å, forces are predicted with good accuracy compared to the force magnitude of QM force, as seen from Fig.8.3.

Summary. In this Chapter, the MLOTF methodology is extended to the calculation of more complex chemical environments such as binary compounds of SiC and SiO_2. With the preliminary results, it becomes clear that the machine-learning capability of the MLOTF methodology can be transferable.

Chapter 9

Conclusion

In this Thesis work, I proposed a novel scheme for Machine Learning of Quantum Mechanically computed atomistic forces, where the inference procedures necessary for force prediction are carried out in the framework of Gaussian Process regression. Great emphasis has been put on force accuracy, and how to handle the information available for force prediction as stored in a dynamically updated configuration database. A key factor for efficiently applying the ML technique was the definition of a covariance matrix between database configurations, which made it necessary to design, implement, and validate a new, vector-based configuration representation. This captures the relevant features of a given atomic environment via a group of internal vectors $\{v_i\}(i = 1, 2, \cdots, k)$ having by construction the same symmetry of the target QM force vector. Any difficulty related to the challenging issue of reference frame dependence could be altogether avoided by associating, for every configuration/internal-vector set a rotationally invariant matrix of mutual projections of the internal vectors onto each other, and constructing an appropriately conditioned metric to measure the distance between any such matrix pair, thus defining the database topology for force prediction.

An advantage of using this two-level representation is that the internal vector set constructed from the atomic positions can be augmented by addition of further vector quantities which are deemed meaningful for QM-accurate force prediction. These might be, e.g., atomic forces calculated using a classical force field or an empirical QM Hamiltonian. Interestingly, while these forces may at any time deviate from the DFT-level forces

by significantly more than the tolerance which the present method aims for, the systematic way these forces correlate with the DFT target one can be efficiently "learned" in the form of the improved database topology associated with the augmented internal vector set. Actual testing shows that including classical and empirical QM forces in the way just described greatly favours the accuracy of force prediction, getting remarkably close to the QM benchmark values (a rather tight force tolerance of the order of ~ 0.1 eV/Å is the general target of the method).

The current scheme for machine-learning-on-the-fly ('MLOTF') differs from ML potentials -whether represented as Gaussian Approximated Potentials (GAP) or Neural Networks- which have recently been proposed to machine-learn the system's potential energy surface 'once-and-for-all', in that it is targeted at enhancing the standard predictor-corrector ability of interpolating accurate QM forces during large-scale molecular dynamics simulations. This guarantees that the desired average accuracy is achieved at all times, while no 'atomic energy' or total energy expression is ever required. At the same time, the option is kept open to develop new information whenever a novel chemical situation is encountered along the system trajectory which necessitates database augmentation though novel QM calculations.

Ideally, the configuration database is updated only when such chemically novel configurations are encountered, and to the extent that this is achieved, ideal information efficiency and (connectedly) large acceleration factors over standard reference first-principles MD can be achieved (e.g., a factor of ~ 30 or more for Silicon at 1000 K). A central result of this thesis work is that the finally produced practical implementation of the method significantly improves on the previous 'Learn-On-The-Fly' (LOTF) molecular dynamics scheme, which was a purely predictor-corrector one and thus made no attempt of storing and re-using the valuable QM information computed at the predictor stage. A further finding is that databases generated by MLOTF simulations are transferable to different simulation runs, so that databases build up along projects. At the same time, an importance sampling criterion whereby only the closest $N_{teach} \sim 500$ configurations are ever used for force prediction keeps the prediction stage optimally fast in production calculations. The overall MLOTF methodology resulting from this work looks very promising for use in multi-scale simulations where a large embedding portion of the system typically hosts very little new chemical activity during the simulations, while the chemically active embedded region(s) is (are) highly localised (e.g., to the atomic region near the

crack tip during fracture propagation by iterative bond breaking). For problems of this kind a relatively small database is typically sufficient to describe well the chemically inactive part, while ever finer/larger databases are developed during MLOTF dynamics which are able to capture the subtle processes happening in the chemically interesting area.

Investigating whether in situations like this the QM-evaluation/force learning can ever be switched off at all times took the present work into extensive testing of a setup where a chemically active crack tip region described by a GAP-class potential was embedded in a larger brittle matrix described by an off-the-shelf Stillinger-Weber classical force field. While no strict accuracy claim could be maintained in this part of the work (which to some extent confirms that MLOTF-class open learning approaches are necessary for ultimate accuracy), this line of investigation produced a useful qualitative physical picture of how the crack speed can be expected to change upon varying the temperatures and loading rates emerged from this work. This was rationalised on the basis of a very simple model relating crack tip bond breaking process to the (temperature dependent) initial population of local vibrational modes, the frequencies of such modes, and the crack propagation speed.

The extending of the application of MLOTF to multi-species was carried out on binary prototypes, SiC and SiO_2. The represented atomic environments are made up of internal vectors derived from sub-lattices of each species and the cross projections of these two groups of internal vectors. The prediction on the preliminary test calculations already shows good accuracy compared to the classical description and is encouraging sign for moving to more challenging chemical situations.

Chapter 10

Outlook of MLOTF

1. The internal-vector representation discussed in this thesis generally works for the QM-force related atomic environments. Regarding the optimisation procedure for generating the representing vectors, as opposed to the empirical selection of a vector subset, it is possible to carry it out satisfying the requirements of (a) correlation with the target QM force; (b) completeness in representing the atomic environments. As for the multi-species system, the internal vector representation expands in the complexity spectrum while the optimisation becomes key to maintaining accuracy and efficiency, ideally with the permutational symmetries in the system taken into account. An optimal representation only captures the pertinent structural variation. Following this reasoning, the optimisation can be dynamically performed for each test configuration, only in the domain of the structural changes we are concerned with. The proposed multi-species representation scheme using sub-lattice can be optimised to its reduced form and generalised to incorporate any number of chemical species.

2. A rigorous system for error prediction would be enormously useful for guiding the ML calculation to the maximum efficiency. As noted in the main text, the derived variances from Gaussian Process are not strongly correlated with the real error and therefore a rigorous error prediction procedure would greatly enhance the applicability of the methodology into machine learning prediction with better efficiency.

3. When it comes to MLOTF prediction using a QM database with a size of the order of magnitude of 10^6, the computational cost becomes prohibitive even with the $O(N)$ scaling. Application of the methodology requires sparsification of the database together with construction and implementation of a hierarchical ordering of the distances within the database. This data-based methodology can be advanced with more sophisticated machine-learning algorithms.

4. It would be scientifically interesting to further explore the application for use in modelling complex chemical environments, i.e. surfaces, defects, amorphous materials and also to investigate more challenging material behaviours (structural phase transition, for example, analysing the melting curve of a solid). These advanced applications of the developed methodology could be its ultimate goal.

Appendix A

Appendix A

(I) **FGP.F95** : Force Gaussian Processes (FGP) is the Fortran code that I implemented the Gaussian Processes machine-learning of the Hellmann-Feynman forces together with represented atomic environments. The main program is designated for performing the GP prediction upon atomic structures, and making force prediction for any atomic environments under the internal-vector representation.

There are a couple of points to note when deriving the internal-vector representation. The representation is calculated based on weight functions with parameters (r_0, m) from 'GRID.DAT' file. A group of independent vectors are generated with the different pairs of (r_0, m). The representing vectors can be selected to describe the local atomic environments, for instance, some conditions to diminish the overlapping of representation vectors as much as possible. Two of the conditions are:

$$\frac{\partial|\vec{V}|}{\partial m} > \delta$$

and

$$\frac{\partial|\vec{V}|}{\partial r_0} > \delta$$

where, δ is increment precision for the variation of the internal vectors with respect to the pair of built-in weight parameters, r_0 and m. Under the conditions above, only the vectors with significant variation with respect to the pair of parameters would be incorporated into the representation. Other optimisation procedures include considering

the dependence relation between each pair of internal vectors and correlation with the target QM forces.

(II) **The Code can also Perform Tasks as follows :**,

1. Abstracting ML information from given teaching configurations, calculating the internal Vectors and building the covariance matrix for the database. All the data (including the Internal Vectors, and QM forces) are stored into the teaching information

2. To construct the covariance matrix:

   A. If performing sorting /selecting algorithms, the sub-database containing the closest teaching configurations will be constructed and the covariance matrix is calculated upon.

   B. constructing the covariance between the test configurations and configurations in the teaching database.

3. To perform the Gaussian Process prediction for the force components on each of the internal directions, with the results including both mean predictive value and GP variance for each predicted components

4. Based on the predicted force components to calculate the most-likely force vector that coordinates the individual components and error evaluation for the process are calculated which can be utilised as an indicator for the confidence level associated with the predicted components.

5. If updating the database is provided, the information of the test configuration will be appended to the database.

Bibliography

[1] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 1964.

[2] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965.

[3] Ting Zhu, Ju Li, Xi Lin, and Sidney Yip. Stress-dependent molecular pathways of silica–water reaction. *Journal of the Mechanics and Physics of Solids*, 53(7): 1597–1623, 2005.

[4] Rubén Pérez and Peter Gumbsch. Directional anisotropy in the cleavage fracture of silicon. *Physical Review Letters*, 84(23):5347, 2000.

[5] Furio Ercolessi and James B Adams. Interatomic potentials from first-principles calculations: the force-matching method. *EPL (Europhysics Letters)*, 26(8):583, 1994.

[6] Y. Mishin, D. Farkas, M. Mehl, and D. Papaconstantopoulos. Interatomic potentials for monoatomic metals from experimental data and ab initio calculations. *Phys. Rev. B.*, 59(5):3393–3407, 1999. doi: 10.1103/PhysRevB.59.3393.

[7] Martin Z Bazant, Efthimios Kaxiras, and JF Justo. Environment-dependent interatomic potential for bulk silicon. *Physical Review B*, 56(14):8542, 1997.

[8] P. Tangney and S. Scandolo. An ab initio parametrized interatomic force field for silica. *J. Chem. Phys.*, 117(19):8898, 2002. doi: 10.1063/1.1513312.

[9] Albert P. Bartók, Mike C. Payne, and Gábor Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics without the Electrons. *Phys. Rev. Lett.*, 104(13):1–4, 2010. doi: 10.1103/PhysRevLett.104.136403.

[10] Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

[11] Noam Bernstein, J R Kermode, and G. Csányi. Hybrid atomistic simulation methods for materials systems. *Reports on Progress in Physics*, 72(2):026501, 2009. ISSN 0034-4885. doi: 10.1088/0034-4885/72/2/026501. URL http://stacks.iop.org/0034-4885/72/i=2/a=026501?key=crossref.4bda1635f633b3d7c80eae9293c476c7.

[12] Gabor Csányi, T. Albaret, M. Payne, and A. De Vita. Learn on the Fly: A Hybrid Classical and Quantum-Mechanical Molecular Dynamics Simulation. *Phys. Rev. Lett.*, 93(17):175503, 2004. doi: 10.1103/PhysRevLett.93.175503.

[13] J. R. Kermode, T. Albaret, Dov Sherman, Noam Bernstein, P. Gumbsch, M. C. Payne, G. Csányi, and A. De Vita. Low-speed fracture instabilities in a brittle crystal. *Nature*, 455(7217):1224–1227, 2008. doi: 10.1038/nature07297.

[14] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[15] PR Bunker. The nuclear mass dependence of the dunham coefficients and the breakdown of the Born-Oppenheimer approximation. *Journal of Molecular Spectroscopy*, 68(3):367–371, 1977.

[16] François Gygi and Giulia Galli. Electronic excitations and the compressibility of deuterium. *Physical Review B*, 65(22):220102, 2002.

[17] Alfredo A Correa, Jorge Kohanoff, Emilio Artacho, Daniel Sánchez-Portal, and Alfredo Caro. Nonadiabatic forces in ion-solid interactions: The initial stages of radiation damage. *Physical Review Letters*, 108(21):213201, 2012.

[18] Robert G Parr and Weitao Yang. *Density functional theory of atoms and molecules*, volume 16. Oxford university press, 1989.

[19] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[20] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087, 1953.

[21] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048–5079, May 1981. doi: 10.1103/PhysRevB.23.5048. URL http://link.aps.org/doi/10.1103/PhysRevB.23.5048.

[22] Hendrik J Monkhorst and James D Pack. Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192, 1976.

[23] Mike C Payne, Michael P Teter, Douglas C Allan, TA Arias, and JD Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Reviews of Modern Physics*, 64(4):1045–1097, 1992.

[24] David Vanderbilt. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Physical Review B*, 41(11):7892, 1990.

[25] Georg Kresse and Jürgen Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996.

[26] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16): 11169, 1996.

[27] Zhenwei Li, Ying Xu, Guoying Gao, Tian Cui, and Yanming Ma. Tetragonal high-pressure phase of ZnO predicted from first principles. *Physical Review B*, 79(19): 193201, 2009.

[28] J. Hone, M. Whitney, C. Piskoti, and A. Zettl. Thermal conductivity of single-walled carbon nanotubes. *Phys. Rev. B*, 59:R2514–R2516, Jan 1999. doi: 10.1103/PhysRevB.59.R2514. URL http://link.aps.org/doi/10.1103/PhysRevB.59.R2514.

[29] P. B. Allen and R. C. Dynes. Transition temperature of strong-coupled superconductors reanalyzed. *Phys. Rev. B*, 12:905–922, Aug 1975. doi: 10.1103/PhysRevB.12.905. URL http://link.aps.org/doi/10.1103/PhysRevB.12.905.

[30] K. Parlinski, Z. Q. Li, and Y. Kawazoe. First-Principles Determination of the Soft Mode in Cubic ZrO2. *Phys. Rev. Lett.*, 78(21):4063–4066, 1997. doi: 10.1103/PhysRevLett.78.4063.

[31] John C Slater and George F Koster. Simplified LCAO method for the periodic potential problem. *Physical Review*, 94(6):1498, 1954.

[32] Madhu Menon and K. R. Subbaswamy. Transferable nonorthogonal tight-binding scheme for silicon. *Phys. Rev. B*, 50:11577–11582, Oct 1994. doi: 10.1103/PhysRevB.50.11577. URL http://link.aps.org/doi/10.1103/PhysRevB.50.11577.

[33] Th. Frauenheim, G. Seifert, M. Elsterner, Z. Hajnal, G. Jungnickel, D. Porezag, S. Suhai, and R. Scholz. A self-consistent charge density-functional based tight-binding method for predictive materials simulations in physics, chemistry and biology. *Physica Status Solidi(b)*, 217(1):41–62, 2000.

[34] Ahmad Jabbarzadeh and Roger I Tanner. Molecular dynamics simulation and its application to nano-rheology. *Rheology Reviews*, 2006:165, 2006.

[35] R. G. Palmer. Broken ergodicity. *Advances in Physics*, 31:669–735, November 1982. doi: 10.1080/00018738200101438.

[36] Loup Verlet. Computer experiments on classical fluids. I. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159(1):98, 1967.

[37] Anderson H C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384, 1980.

[38] Herman JC Berendsen, J Pl M Postma, Wilfred F van Gunsteren, ARHJ DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984.

[39] D. Quigley and M.I.J. Probert. Langevin dynamics in constant pressure extended systems. *The Journal of Chemical Physics*, 120:11432, 2004.

[40] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81:511, 1984.

[41] Noam Bernstein, Csilla Várnai, Iván Solt, Steven A Winfield, Mike C Payne, István Simon, Mónika Fuxreiter, and Gábor Csányi. QM/MM simulation of liquid water with an adaptive quantum region. *Physical Chemistry Chemical Physics*, 14 (2):646–656, 2012.

[42] John E Lennard-Jones. Cohesion. *Proceedings of the Physical Society*, 43(5):461, 1931.

[43] Frank H Stillinger and Thomas A Weber. Computer simulation of local order in condensed phases of silicon. *Physical Review B*, 31(8):5262, 1985.

[44] Daniela Kohen, John C Tully, and Frank H Stillinger. Modeling the interaction of hydrogen with silicon surfaces. *Surface Science*, 397(1):225–236, 1998.

[45] Walter A Harrison. *Electronic structure and the properties of solids: the physics of the chemical bond*. Courier Dover Publications, 2012.

[46] J. R. Kermode. *Multiscale Hybrid Simulation of Brittle Fracture*. PhD thesis, Pembroke College, University of Cambridge, 2008.

[47] J. Tersoff. Empirical interatomic potential for silicon with improved elastic properties. *Phys. Rev. B*, 38(14):9902–9905, 1988.

[48] J. Tersoff. New empirical model for the structural properties of silicon. *Physical Review Letters*, 56(6):632–635, 1986.

[49] Donald W Brenner. Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. *Physical Review B*, 42(15):9458, 1990.

[50] Michele Parrinello. From silicon to RNA: The coming of age of ab initio molecular dynamics. *Solid State Communications*, 102(2):107–120, 1997.

[51] Richard Car and Mark Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55(22):2471, 1985.

[52] Chris-Kriton Skylaris, Peter D Haynes, Arash a Mostofi, and Mike C Payne. Introducing ONETEP: linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.*, 122(8):84119, 2005. doi: 10.1063/1.1839852.

[53] José M Soler, Emilio Artacho, Julian D Gale, Alberto García, Javier Junquera, Pablo Ordejón, and Daniel Sánchez-Portal. The SIESTA method for ab initio order-n materials simulation. *Journal of Physics: Condensed Matter*, 14(11):2745, 2002.

[54] A De Vita and R Car. A novel scheme for accurate MD simulations of large systems. *Mater. Res. Soc. Symp. Proc*, 491:473–480, 1998.

[55] Arieh Warshel and Michael Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2):227–249, 1976.

[56] Richard A Friesner and Victor Guallar. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.*, 56:389–427, 2005.

[57] Feliu Maseras and Keiji Morokuma. Imomm: A new integrated ab initio+ molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *Journal of Computational Chemistry*, 16(9):1170–1179, 1995.

[58] A. Laio, J. VandeVondele, and U. Rothlisberger. A hamiltonian electrostatic coupling scheme for hybrid car-parrinello molecular dynamics simulation. *J. Chem. Phys.*, 116(6941), 2002.

[59] James Kermode, Giovanni Peralta, Zhenwei Li, and Alessandro De Vita. Multiscale modelling of materials chemomechanics: brittle fracture of oxides and semiconductors. *Procedia Materials Science*, 3:2681–2686, 2014.

[60] B. Lawn. *Fracture of Brittle Solids*. Cambridge University Press, 1993.

[61] A.A. Griffith. The phenomena of rupture and flow in solids. *Philos. Trans. R. Soc. London A*, 221(163), 1921.

[62] R. Thomson, C. Hsieh, and V. Rana. Lattice Trapping of Fracture Cracks. *Journal of Applied Physics*, 42(3154), 1971.

[63] L. B. Freund. *Dynamic Fracture Mechanics*. Cambridge University Press, 1990.

[64] Jay Fineberg, Steven Gross, M Marder, and Harry Swinney. Instability in dynamic fracture. *Physical Review Letters*, 67(4):457–460, July 1991. ISSN 0031-9007. doi: 10.1103/PhysRevLett.67.457.

[65] J Fineberg and M Marder. Instability in dynamic fracture. *Physics Reports*, 313 (1-2):1–108, May 1999. ISSN 03701573. doi: 10.1016/S0370-1573(98)00085-4.

[66] F. Atrash, A. Hashibon, P. Gumbsch, and D. Sherman. Phonon emission induced dynamic fracture phenomena. *Phys. Rev. Lett.*, 106:085502, Feb 2011. doi: 10.1103/PhysRevLett.106.085502. URL http://link.aps.org/doi/10.1103/PhysRevLett.106.085502.

[67] Dominic Holland and M. Marder. Ideal brittle fracture of silicon studied with molecular dynamics. *Phys. Rev. Lett.*, 80:746–749, Jan 1998. doi: 10.1103/PhysRevLett.80.746. URL http://link.aps.org/doi/10.1103/PhysRevLett.80.746.

[68] Dominic John Martin Holland and Michael Marder. *Cracks and atoms*. University of Texas at Austin, 1999.

[69] N Bernstein and DW Hess. Lattice trapping barriers to brittle fracture. *Physical Review Letters*, 91(2):025501, 2003.

[70] Markus J Buehler. *Atomistic modeling of materials failure*. Springer, 2008.

[71] M. Marder and Xiangming Liu. Instability in lattice fracture. *Phys. Rev. Lett.*, 71 (2417), 1993.

[72] Jens A. Hauch, Dominic Holland, M. P. Marder, and Harry L. Swinney. Dynamic fracture in single crystal silicon. *Phys. Rev. Lett.*, 82(3823), 1999.

[73] Ilan Beery, Uri Lev, and Dov Sherman. On the lower limiting velocity of a dynamic crack in brittle solids. *Journal of Applied Physics*, 93(5):2429–2434, 2003.

[74] Giovanni Peralta, James Kermode, Silvia Cereda, Zhenwei Li, and Alessandro De Vita. Inversion in crack speed versus temperature during crossover from activated to catastropic fracture. *In Preparation*, 2014.

[75] K. C. Pandey. New π-bonded chain model for Si(111)-(2 $\times$ 1) surface. *Phys. Rev. Lett.*, 47(1913), 1981.

[76] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.*, volume 14. MIT Press, 2006. ISBN 026218253X.

[77] David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge University Press, 2003.

[78] Mark Gibbs and David JC MacKay. Efficient implementation of gaussian processes. 1997.

[79] Thomas B Blank, Steven D Brown, August W Calhoun, and Douglas J Doren. Neural network models of potential energy surfaces. *The Journal of Chemical Physics*, 103(10):4129–4137, 1995.

[80] Paul J Steinhardt, David R Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 28(2):784, 1983.

[81] Wolfgang Lechner and Christoph Dellago. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of Chemical Physics*, 129:114707, 2008.

[82] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[83] Albert P Bartók, Michael J Gillan, Frederick R Manby, and Gábor Csányi. Machine-learning approach for one-and two-body corrections to density functional theory: Applications to molecular and condensed water. *Physical Review B*, 88 (5):054104, 2013.

[84] Wojciech J Szlachta, Albert P Bartók, and Gábor Csányi. Accuracy and transferability of gap models for tungsten. *arXiv preprint arXiv:1405.4370*, 2014.

[85] Hagai Eshet, Rustam Z. Khaliullin, Thomas D. Kühne, Jörg Behler, and Michele Parrinello. Microscopic origins of the anomalous melting behavior of sodium under high pressure. *Phys. Rev. Lett.*, 108:115701, Mar 2012. doi: 10.1103/PhysRevLett.108.115701. URL http://link.aps.org/doi/10.1103/PhysRevLett.108.115701.

[86] Hagai Eshet, Rustam Z Khaliullin, Thomas D Kühne, Jörg Behler, and Michele Parrinello. Ab initio quality neural-network potential for sodium. *Physical Review B*, 81(18):184107, 2010.

[87] Rustam Z Khaliullin, Hagai Eshet, Thomas D Kühne, Jörg Behler, and Michele Parrinello. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nature Materials*, 10(9):693–697, 2011.

[88] Nongnuch Artrith, Tobias Morawietz, and Jörg Behler. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B*, 83:153101, Apr 2011. doi: 10.1103/PhysRevB.83.153101. URL http://link.aps.org/doi/10.1103/PhysRevB.83.153101.

[89] Tobias Morawietz, Vikas Sharma, and Jörg Behler. A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *The Journal of Chemical Physics*, 136(6):064103, 2012.

[90] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012.

[91] John C Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding density functionals with machine learning. *Physical Review Letters*, 108(25):253002, 2012.

[92] G. Moras G. Csanyi A. Peguiron, J.R. Kermode and A. De Vita. Accuracy of QM/MM electrostatic embedding schemes in silica. *In Preparation*, 2014.

[93] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

[94] Alan L Mackay. Quaternion transformation of molecular orientation. *Acta Crystallographica Section A: Foundations of Crystallography*, 40(2):165–166, 1984.

[95] Artem R Oganov, Yanming Ma, Andriy O Lyakhov, Mario Valle, and Carlo Gatti. Evolutionary crystal structure prediction as a method for the discovery of minerals and materials. *Reviews in Mineralogy and Geochemistry*, 71(1):271–298, 2010.

[96] John B Carroll. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26(4):347–372, 1961.

[97] Wim Meeusen and Julien Van den Broeck. Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review*, 18(2): 435–444, 1977.

[98] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

[99] Atsushi Togo, Fumiyasu Oba, and Isao Tanaka. First-principles calculations of the ferroelastic transition between rutile-type and CaCl2-type SiO_2 at high pressures. *Phys. Rev. B.*, 78(13):1–9, 2008. doi: 10.1103/PhysRevB.78.134106.

[100] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press, 2007.

[101] Meijie Tang and Sidney Yip. Lattice instability in β-SiC and simulation of brittle fracture. *Journal of Applied Physics*, 76(5):2719–2725, 1994.

[102] Meijie Tang and Sidney Yip. Atomistic simulation of thermomechanical properties of β-SiC. *Physical Review B*, 52(21):15150, 1995.