THE UNIVERSITY *of* EDINBURGH

Developing Methods to Machine-Learn Potentials with application to Nitrogen

Marcin Kirsz

Doctor of Philosophy
The University of Edinburgh
December 2023

Abstract

Computational studies of condensed matter phases by molecular dynamics are limited by the lack of accurate and efficient interatomic potentials. The high level theories, such as density functional theory (DFT), provide accurate potential energy surface description but lack required efficiency for large scale problems. On the other end of the spectrum are empirical potentials which are fast but often not accurate enough. The emergence of new machine learning methods for the development of interatomic potentials aim to bridge this gap.

This thesis presents the development of machine learning library for interatomic potentials. The *Ta-dah!* software is capable of generating machine-learned potentials for mono- and multi-component systems. The library provides wide range of atomic local environment descriptors and its modular structure allows quick implementation of new ideas. The library is fully interfaced with LAMMPS molecular dynamics software.

The standard use of *Ta-dah!* involves training with data generated from DFT packages such as VASP and CASTEP. It also incorporates a training method for learning interatomic potentials from high level quantum mechanical theories, such as coupled cluster. The method allows to harvest existing databases of high quality quantum chemistry calculations to build interatomic potentials based on methods which, in principle, can exceed that achievable by density functional theory.

The library is deployed to develop efficient and accurate interatomic potentials to study various systems. The thesis highlights molecular dynamics calculations with a new potential for molecular nitrogen, based on quantum chemistry data. Phase-coexistence and free energy calculations with this potential are used to describe the melt curve and several different crystal phases. This enables calculation of the phase diagram up to 10 GPa. The potential is also applied

in the to study of the proposed “Frenkel Line” in the subcritical and supercritical regions.

Lay Summary

The physical sciences have progressed through a meticulous process of developing concepts and verifying them with experiment. Regardless of how beautiful the theory is, if it disagrees with the experiment it has to be refined or even replaced by new improved model. In short, the experiment is king.

However, over the past few decades the computational methods gained significant traction and nowadays are considered as an indispensable tool in science. They are not only used to complement experiments but perhaps more importantly they can guide them towards new discoveries and in consequence significantly accelerate progress of science. Moreover, the computational methods allow us to study phenomena inaccessible to experiment, such as what happens in the interior of the Earth or even remote objects such as newly discovered exoplanets found beyond the Solar System.

Of particular importance are computational methods which let us simulate matter on a microscopic level. They allow us to model individual atoms and track them on their journey through space and time. The quality of those simulations and the motion of atoms itself is governed by the underlying theoretical model of the forces they exert on one another - the interatomic potential.

The great deal of physics is concerned with the development of better and more accurate interatomic potentials. While computational methods based on the laws of quantum mechanics are extremely accurate, they lack the computational efficiency to study more than a few atoms. There is an increasing need for better interatomic potentials capable of simulating large systems with a high level of accuracy.

In this work a data-driven approach is used to transcribe complicated quantum mechanical laws into efficient interatomic potentials which allow simulations for even millions of atoms. The new method uses well established machine

learning algorithms which are becoming increasingly popular in almost all aspects of science. The software developed for this PhD thesis permits *training* of interatomic potentials which then can be used to study solids and liquids on the microscopic level.

The software is used to develop an interatomic potential for nitrogen. The model is extremely accurate and matches available experimental data very well. The new interatomic potential is relevant to study nitrogen solids and liquids in a wide range of temperatures and pressures. The model can be used to study nitrogen ices found on many planets which are not directly accessible to the experiment. For example, it is suitable for investigation of icy nitrogen fields found on planet Pluto by the New Horizon spacecraft. This will deepen our understanding on the formation of this system as well as shed some light on the transformation of the early Solar System.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in [143, 144].

(*Marcin Kirsz, December 2023*)

Acknowledgements

I thank my supervisor, Graeme Ackland, for providing me with the limitless support, trust and believe that I can make it. For hours of discussions and plethora of valuable insights which helped to shape me as a physicist.

I thank Hongxiang Zong for introducing me to the field of machine learning interatomic potentials, his kindness and enthusiasm.

I thank Andreas Hermann, Andrew Huxley and Miguel Martinez-Canales for helping me to stay on track, sharing their knowledge and a kind word when it was needed.

I thank Paul Clegg for his support when things were not going that well.

I thank Ciprian Pruteanu for fruitful collaboration and a friendly nudge when it came to publishing our work.

I thank Gavin Woolman, Asuka Nakamura-Pinder and Elspeth Smith for giving *Ta-dah!* a go, even though the code was far from polished at the time.

Last but not least I thank Magdalena, Iris and Teo for their support, motivation and love. I could not have undertaken this journey without you - and no, you do not have to read this thesis now :-)

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ACSF	Atom Centred Symmetry Function
API	Application Programming Interface
bcc	Body-Centred Cubic
BLR	Bayesian Linear Regression
BO	Bayesian Optimisation
CC	Coupled Cluster
CCS	Coupled Cluster Single
CCSD	Coupled Cluster Single Double
CCSD(T)	Coupled Cluster Single Double with perturbative Triple
CIP	Classical Interatomic Potential
CLI	Command Line Interface
CP	Critical Point
DFT	Density Functional Theory
DM	Design Matrix
EA	Evolutionary Algorithm
EAD	Embedded Atom Descriptor
EKM	Empirical Kernel Map
EoS	Equation of State
fcc	Face-Centred Cubic
FL	Frenkel Line
GA	Genetic Algorithm
GAP	Gaussian Approximation Potential
GD	Gradient Descent
GLF	Global Loss Function
GOA	Global Optimisation Algorithm
GP	Gaussian Process
GPR	Gaussian Process Regression
hcp	Hexagonal Close Packed
HP	Hyper Parameter
IP	Interatomic Potential
KRR	Kernel Ridge Regression
LP	Learned Parameters
LR	Linear Regression
MAE	Mean Absolute Error
MD	Molecular Dynamics

MLIP	Machine Learning Interatomic Potential
NIST	National Institute of Standards and Technology
NN	Neural Network
PC	Performance Constraint
PES	Potential Energy Surface
PSO	Particle Swarm Optimisation
RDF	Radial Distribution Function
RMSE	Root Mean Square Error
SSC	Search Space Constraint
VACF	Velocity Autocorrelation Function
WL	Widom Line
QM	Quantum Mechanics
5CM	Five-centre site-site model

Chapter 1

Introduction

Advances in our understanding of the Universe happened as a consequence of human curiosity, intuition, laborious trial and error process and, perhaps mostly, a fortunate stroke of serendipity. Few of those are within our control. However, with the accumulated knowledge, the trial and error process can be refined and successfully applied to progress our understanding of the Laws of Nature.

Since the early days of physics, the experiment was, and still remains, the final judge of scientific truth [62]. In the second half of the 20th century, atomic simulations have become a common tool to aid our understanding of Nature in fields of physics, chemistry and material science. Computer simulations are not only used to shed light on the experimental data but also to make predictions which can later be verified by the experiment. The ability to study problems on the microscopic level allows researchers to gain fundamental insight into the nature of underlying processes.

For centuries, discoveries of new materials have been dominated by the trial and error process. The demand for materials with specific properties is higher than ever before and can only continue to grow. Either the fine-tuning of material properties to match specific application, or a discovery of completely new materials with remarkable and sometime exotic properties are very time consuming. Nowadays, this process of material design and testing can be greatly accelerated with the use of computer simulations. In consequence, modern material science is being transformed by novel computer modelling techniques [72].

Atomic simulations is also a standard instrument used in planetary science even though the time and length scales can span many orders of magnitude. Understanding of the relevant processes, such as phase transitions, is crucial to address questions such as dynamics of the Earth interior. This high temperature and pressure environment is often inaccessible to the experiment but can be studied at the molecular level with simulations such as Molecular Dynamics (MD) or Monte Carlo (MC). Moreover, the humanity is turning its attention not just beyond Earth but even beyond the Solar System. With the discoveries of new exoplanets it is only natural to turn towards computational methods, such as atomic simulations, to study matter in the conditions found on those remote objects.

The atomic simulations critically depend on the choice of the interatomic potential [182]. The role of the potential is to accurately describe interactions between atoms such that their collective behaviour give rise to the phenomena later observed in Nature. The choice of the potential is dictated by the physical question in mind and limited by the available computational resources. In general, accurate potentials are computationally more demanding. Henceforth there is a increasing need for more accurate interatomic potentials which are able to simulate systems with a large number of atoms [127].

Historically the interatomic potentials are broadly divided into two groups: Ab initio and empirical. Ab initio (also known as "from first principles") are potentials where many body Schrödinger equation is solved computationally to obtain accurate energies from which forces can be calculated thanks to the Hellmann-Feynman theorem [61]. Arguably the Density Functional Theory (DFT) [88, 104] is the most successful theory to date which allows us to predict energies on the milli electronvolt scale. The most accurate energy predictions are obtainable from quantum chemical wavefunction based theories such as Coupled Cluster (CC) [40, 41, 43, 44] however at great computational expense.

The major limiting factor of first principle methods is their computational scalability which is $\mathcal{O}(n^3)$ with the number of atoms for the DFT. Even with an exponential growth in computing power and increased availability of high performance computing facilities for many scientific groups world-wide the ab initio methods are still unable to address time and length scales required to tackle many interesting problems in physics. In practice, DFT can simulate systems of hundreds of particles on the picosecond scale. The CC method is limited to computing interactions between simple molecules when highly accurate energies

are required.

On the other end of the spectrum are empirical potentials which aim to reproduce quantum mechanical effects using functional form motivated by the higher theory. These potentials contains free parameters such as bond lengths and angles, charge, etc., which are then fitted to either empirical data (e.g. lattice parameters, elastic constants) or to quantum mechanical data from ab initio theory. The empirical potentials allow systems of hundreds of thousands of atoms to be simulated on the nanosecond scale however they lack accuracy and their capabilities are limited to often narrow domain of applicability.

In the early days of empirical potentials, their functional forms where mathematically simple and had clear physical meaning. Over time, the functional forms evolved from simple functions towards more complex nonlinear models. Still, the resolution of forces and energies from the empirical potentials is often not sufficient for many problems in physics.

Number of different empirical potentials have been developed since 1980s. Notable many-body potentials to describe metallic systems are Embedded Atom Method (EAM) [47, 48] and Finnis-Sinclair type model [64] while covalent materials are often simulated with bond-order type potentials [174, 184–186]. New functional forms are still being developed to better describe the chemical bonding found across various classes of materials [127]. It is worth emphasising that the construction of empirical potentials requires extended domain knowledge and is often considered as an art and science in equal parts. Some high quality potentials are used extensively by the community [1–3, 123, 124] while some others, perhaps of lesser quality, are never being used after initial publication.

The interatomic potential's ability to represent accurately atomistic phenomena is by far the most important factor which determines its popularity. However, interatomic potentials which are implemented in easy-to-use software are more likely to be employed in the future. The platforms like National Institute of Standards and Technology (NIST) Interatomic Potentials Repository [15, 80] and the Knowledgebase of Interatomic Models (OpenKIM) [181] provide thousands of validated and ready-to-use potentials. Moreover, the distributed models seamlessly work with many popular MD packages such as Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [189] either via Application Programming Interface (API) or the file format which is natively supported by the MD package. The testing and verification of the models in those

databases assist prospective users in decision whether the potential is applicable to the problem at hand.

There is an increasing need for efficient potentials which can simulate tens (or even hundreds) of thousands of atoms on the nanosecond scale with the accuracy of quantum mechanical codes. It has been considered for a while now that the methods employed in the development of empirical potentials have reached a 'plateau' and therefore are insufficient in terms of accuracy and reliability for more advanced applications [72]. The main limiting factor is a functional form of the interatomic potential which is in general unknown and must be guessed. Clearly new directions were needed. At the same time the interdisciplinary data-driven approaches have become increasingly popular in many fields of science (see for example [6, 83, 212]). The new paradigm emerged utilising those ideas. In particular, the developments in machine learning (ML) methods allow us now to address the increasing need for efficient interatomic potentials which can provide accurate energies and forces.

The roots of machine learning can be traced back to the early 19th century with the discovery of least squares method by Adrien-Marie Legendre in 1805 and rediscovery of Bayes Theorem by Pierre-Simon Laplace seven years later. However, it was not until 1950 when Alan Turing proposed a "learning machine" that can be trained to become "intelligent" [194]. The first pioneering attempt in the field of machine learning was construction of SNARC (Stochastic Neural Analog Reinforcement Calculator) by Marvin Minsky in 1951. SNARC was imitating a neural net inspired by the biological neural networks with approximately 40 randomly connected synapses [122]. Despite the initial success of SNARC, in 1969 the same author published work that shows some limitations of neural network (NN) machines which were widely (and wrongly) interpreted as having a fundamental flaw. The rediscovery of a backpropagation algorithm in the 1980s proved to be a turning point: it meant that a NN can efficiently be trained in an iterative and recursive manner which previously thought is impossible. From the 1990s machine learning methods become widespread in scientific community where are mostly used to analyse large sets of data and "learn" from results. From year 2000 onwards there is a rapid development in machine learning techniques mostly driven by the commercial needs. The probabilistic perspective to ML gave rise to methods such as Support Vector Machines and range of Kernel Methods [26, 131, 148].

Machine learning techniques have been used in fields of physics and chemistry for

over 30 years [70, 178]. Initially they were employed to help with classification problems (such as the analysis of spectra [191]) but nowadays they assist in wide range of tasks such as identifying phases and phase transitions in variety of condensed matter systems [36] or to efficiently represent quantum many body state using deep Neural Networks (dNN) [67].

Machine learning methods have also emerged as a promising tool for the development of the interatomic potentials [11, 19, 53, 119, 162, 188, 215]. The new paradigm attempts to represent the potential energy surface (PES) by numerical interpolation of reference data generated with higher order quantum mechanical calculations. This is achieved by replacing the physically motivated functional form of empirical potentials with more flexible mathematical one, albeit lacking physical meaning [17].

The key idea in the development of the machine learning interatomic potentials (MLIP) is to construct a functional relation between atomic environments and energy [10]. The potential is then trained using a database from higher levels of theory such as electronic structure calculations. In this way the hierarchy of potentials is obtained where, in principle, quantum mechanical accuracy can smoothly be traded against the efficiency. This allows development of bespoke potentials which are suitable for the wide range of problems [50, 151, 195, 216].

The essential question is what makes a good machine learned interatomic potential? Behler suggests the following requirements [17]:

- Computational cost of evaluation comparable with empirical interatomic potentials.
- The total cost of generating training database should be significantly lower than the direct application of DFT.
- Minimum human effort and intervention should be required during potential training.
- Systematic improvements should be possible.
- Analytic energy gradients should be available for the calculation of forces, stress tensor, etc.
- Potential should not contain any classification of atoms, predefined bonds as chemical environments and bonds should be allowed to change in the course of the simulation.

- High accuracy comparable with the reference electronic calculations.

Another highly sought property of interatomic potentials is their transferability to unknown atomic environments. In general, machine learning interatomic potentials are highly capable in their interpolation ability and mediocre in extrapolation. Therefore this requirement is particularly difficult to satisfy for MLIPs. It has been shown that the problem can be alleviated to some extent by introducing information about the nature of interatomic bonding [114, 146].

The functional forms of MLIPs can vary greatly. Furthermore, the fitting procedures used for training can be very elaborate and often require the development of a separate software library. Moreover, features which may be critical for some applications can be irrelevant for others. It is understandable that no single standard exists to facilitate different codes at this moment in time and it is unlikely that this will happen in the future. It is therefore important that the written code follows well established practices for scientific software development [109, 203, 204]. Simply put, any code which just produces a publication but is not being reused is rarely worthwhile. As the bare minimum the distributed MLIPs should provide an API such that it can be used with a MD package of choice (such as LAMMPS [189]).

The training process for MLIPs aims to minimise the cost function such that the model accurately represents the potential energy surface (PES) for the problem at hand. The concept of PES is related to the Born-Oppenheimer approximation [31] where the motion of atomic nuclei and electrons are separated. That is, given atomic configuration, nuclear and total charge, the potential energy of the system is defined by its electronic Hamiltonian. The MLIP attempts to model this Born-Oppenheimer surface but without calculating the electronic wavefunction explicitly. In fact what we strive to do is to solve an inverse problem: Can the machine "learn" to predict the potential energy surface, given a set of atomic coordinates? The training data are usually obtained from the expensive first principles electronic structure calculations on small systems. The model learns to replicate those energies by capturing the relationship between atomic coordinates and energy using a flexible mathematical set of functions known as descriptors.

The deployment of MLIP in the MD environment involves several steps: transformation of atomic coordinates to a set of feature vectors which enforce fundamental symmetries such as translation, rotation and permutation of indistinguishable atoms. These feature vectors are then used as an input for

the machine learning method. The ML engine is then used to predict energies and forces which are in turn communicated back to the MD code. The atomic positions are then integrated based on the provided forces and new atomic configuration is obtained. Evidently, there is an overlap between training and deployment of the model.

A number of excellent software packages has been published to assist development of MLIPs along with APIs for major MD codes [11, 134, 200]. Those codes are usually fixed by design to a particular method of choice and are in general difficult to extend to accommodate new advances in the field without extensive knowledge of the code base. While this is not a deficiency per se, one might argue that the lack of flexible toolbox for the development of MLIPs hinders progress in the field. It also renders comparison of accuracy and efficiency between different methods difficult [216]. In the last couple of years a number of packages emerged attempting to unify development and deployment of MLIPs [201, 207]. Also, LAMMPS now provides new MLIAP package - an interface which supports some MLIPs albeit still fairly limited in its functionality.

The field of interatomic potentials is dynamic with new methods constantly being developed. There is an increasing need for a toolbox, which would allow one to progress through key stages: development and easy implementation of new methods, training of the model and its efficient deployment to a large scale real-world problems. Arguably, no package developed to this date allows to progress seemingly through all those stages.

The focus of this thesis is on the development of novel software and library for machine learning interatomic potentials for condensed matter systems. The *Ta-dah!* strives to provide an easy to use command line interface (CLI) for training as well as it has ability to be used as a C++ library. The code follows well established software development practices including extensive documentation. The modular structure of the code allows implementation of new ideas with the minimal knowledge of the code base. The deployment of new models is possible by the universal and efficient LAMMPS API. The code features a unique ability to train and deploy simple models for dimers with an accuracy comparable only to Coupled Cluster method as well as novel two-stage fitting procedure.

In this thesis, the *Ta-dah!* software is utilised to develop transferable MLIP for molecular nitrogen. The model is first employed to study the Frenkel Line (FL) in the subcritical and supercritical nitrogen. The improved model is then developed

to compute the phase diagram of N_2 up to 10 GPa.

Further introduction to the nitrogen phase diagram and the Frenkel Line have been incorporated into relevant chapters.

The structure of the thesis is as follows. In chapter 2 the background theory is presented. The major developments in quantum mechanical modelling are summarised in section 2.1. The introduction to classical interatomic potentials, including review of the most popular ones, can be found in section 2.2. The brief overview of molecular dynamics is given in section 2.3. Section 2.4 contains the introduction to machine learning followed by the review of machine learning interatomic potentials in section 2.5.

Chapter 3 presents open-sourced *Ta-dah!* package and its capabilities. The essential software components are discussed in section 3.1 along with an overview of training and prediction procedures. The necessary theory is covered in section 3.3. The chapter ends with a simple example to illustrate *Ta-dah!* usage (section 3.4). The future development directions are listed in section 3.5.

The *Ta-dah!* software features a unique two-stage fitting procedure which is presented in chapter 4. The purpose of it is to improve MLIPs transferability beyond the initial training configurational space. The necessary background is provided in section 4.1 followed by the review of hyper parameter (HP) optimisation techniques in section 4.2. The methodology is laid out in section 4.3 and the proof of the concept is given to validate this approach in section 4.4.

In chapter 5 a new machine learning interatomic potential is developed to study dynamic transition in the subcritical and supercritical nitrogen. The line on the pressure-temperature phase diagram where this transition occurs has been coined the Frenkel Line. The chapter begins with the relevant introductory material followed by the development of the model in section 5.3. The model is validated against the available neutron scattering data and then deployed to study origin of the Frenkel Line on the phase diagram in section 5.4.

Given the excellent agreement between the model and the experiment in liquid simulations, the model is further deployed to compute the phase diagram of nitrogen in chapter 6. We will see that it was found that the initial model is insufficient to reproduce the melt curve, hence the phase diagram of nitrogen. To address shortcoming of the first model, the new, improved methodology and model are developed in section 6.2. The new method allows to train highly

accurate models for dimer-like systems using accurate Coupled Cluster data. The review of relevant experimentally determined solid phases of molecular nitrogen on the phase diagram up to 10 GPa is given in section 6.3. Finally, the phase diagram is computed in section 6.4.

Chapter 2

Background Theory

2.1 Quantum Mechanical Modelling

The behaviour of matter at a macroscopic level is well described by the laws of quantum mechanics. It is postulated that the system of interest can be represented by the wave function Ψ and its time evolution is governed by the time dependent Schrödinger equation (TDSE) given by

$$i\hbar\frac{\mathrm{d}}{\mathrm{d}t}\left|\Psi(t)\right\rangle = \hat{\boldsymbol{H}}\left|\Psi(t)\right\rangle \tag{2.1}$$

where i is the imaginary unit, $\hbar$ is the reduced Planck constant, $|\Psi(t)\rangle$ represents the state of the system at time t and $\hat{\boldsymbol{H}}$ is the Hamiltonian operator which corresponds to the total energy of the system.

The closed form solution is known only for the simplest problems such as isolated hydrogen atom. As soon as additional electrons are introduced, the system becomes to complex to tackle analytically. To address more challenging problems one is required to use approximate methods.

This section provides an overview of the most successful models used to model quantum systems along with the underlying approximations.

2.1.0.1 Born-Oppenheimer Approximation

The BOA [31] has been introduced in 1927 and proved to be a cornerstone for the field of atomistic modelling. It exploits the fact that the rest mass of nuclei is much heavier than of surrounding electrons. Therefore it follows that electrons will respond almost instantaneously to nuclear movement to find their ground state. Based on this, one can decouple motion of electrons from nuclei and write $\hat{\boldsymbol{H}}$ as a product of electronic and nuclear terms. In other words, the motion of nuclei is treated classically while the time independent Schrödinger equation (TISE) is used to solve an electronic wavefunction.

$$\hat{\boldsymbol{H}} \left|\Psi\right\rangle = E \left|\Psi\right\rangle \tag{2.2}$$

The BOA introduces small errors into calculation which can often be neglected. Special attention must be paid when simulating hydrogen molecules as the mass difference from the electron is smaller and the BOA may affect vibrational frequency and at high pressure nuclei motion can no longer be treated classically. There are also other cases where error is significant enough to invalidate results [182]. These may involve light nuclei and systems where nuclei move extremely fast such as during rapid fracture.

2.1.0.2 Variational Method

The variational method allows finding solution to the Schrödinger equation by solving a minimisation problem in the iterative manner which is suitable for modern computers.

The variational theorem states that the true wave function is that which minimises the energy of the system. The wave function can be expanded in a complete orthogonal set of basis functions. By varying parameters of this expansion one can, in principle, converge to the true ground state.

In practice, the approximate wave function is build out of an incomplete set of basis functions (discussed in section 2.1.3.5) and expansion is truncated to balance accuracy and computational effort required. Moreover, the iterative methods used to vary expansion parameters do not guarantee finding the true minima hence obtained energy is an upper bond to the true one.

2.1.1 Hartree

The Hartree method builds on BOA and attempts to solve time independent Schrödinger equation (eq. 2.2) by assuming a system of non interacting electrons. Here the problem is simplified to interaction of individual electrons with all other electrons in the mean-field approximation.

It is a rather bold assumption and cannot be really justified, nevertheless it provides a good starting point for more sophisticated methods.

The total wave function is just a product of individual electron wave functions (orbitals) - known as Hartree product.

$$\Psi(\{\mathbf{r}\}) = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\phi_3(\mathbf{r}_3)...\phi_N(\mathbf{r}_N) \tag{2.3}$$

Apart from neglecting electronic correlation there is another major shortcoming. The Hartree product does not satisfy the anti-symmetric principle which is required for fermions.

Since Hartree electrons do not interact, the Hamiltonian is separable and the goal is to solve TISE for each electronic orbital ϕ

$$(\hat{\boldsymbol{T}}_e + \hat{\boldsymbol{U}}_{Ne} + \hat{\boldsymbol{U}}_{mean})\phi_i(\mathbf{r}_i) = \epsilon_i\phi_i(\mathbf{r}_i) \tag{2.4}$$

where $\hat{\boldsymbol{T}}_e$ is a kinetic energy operator, $\hat{\boldsymbol{U}}_{Ne}$ is a potential due to the fixed nuclei and $\hat{\boldsymbol{U}}_{mean}$ is a mean field due to all the electrons.

The problem is usually solved in the iterative manner based on the variational method (discussed in 2.1.0.2).

2.1.2 Hartree-Fock

The Slater determinant of a set of single electron orbitals has been shown to satisfy the anti-symmetry property required for fermions [51, 84, 166]. This form satisfies quantum mechanical indistinguishability such that each electron is associated

with every orbital

$$\Phi(\{\mathbf{r}\}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r_1}) & \phi_2(\mathbf{r_1}) & \cdots & \phi_N(\mathbf{r_1}) \\ \phi_1(\mathbf{r_2}) & \phi_2(\mathbf{r_2}) & \cdots & \phi_N(\mathbf{r_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{r_N}) & \phi_2(\mathbf{r_N}) & \cdots & \phi_N(\mathbf{r_N}) \end{vmatrix} \tag{2.5}$$

This form allows one to recast a problem into a set of independently moving electrons (orbitals) each experiencing average field from all other electrons and, most importantly, the exchange interaction is properly accounted for.

The resulting Hamiltonian contains additional exchange term on top of what is found in the Hartree method. The term is expensive to compute as it involves integral over whole space.

The HF method is essentially a search of the best set of orbitals which minimises the energy of the corresponding Slater determinant. As in any minimisation problem the resultant energy is an upper bound to the true ground state.

2.1.3 Density Functional Theory

In this section an overview of the method is presented along with key components and fundamental theorems which validate this approach.

The density functional theory (DFT) is a workhorse in computational atomic modelling field. It reformulates SE in terms of electron density, consequently reducing complexity of the problem from 3N dimensions to just 3. In the Kohn Sham approach, the original system is replaced with a fictitious one of N non-interacting electrons experiencing an external effective potential. All the electron-electron interactions are are buried in so so called exchange-correlation term. The aim is to find the ground state density which corresponds to the ground state wavefunction

$$\rho_0(\mathbf{r}) = \langle \Psi_0 | \, \hat{\rho}(\mathbf{r}) \, | \Psi_0 \rangle \tag{2.6}$$

where the density operator at position r is obtained by summing over all electrons denoted by i

$$\hat{\rho}(\mathbf{r}) = \sum_i \delta(\mathbf{r} - \mathbf{r}_i) \tag{2.7}$$

2.1.3.1 Hohenberg–Kohn Theorems

The electron density has been used as a basic variable by Thomas and Fermi in their models published in 1927 [60, 187]. However it was not until 1964 when usage of density has been put onto firm theoretical grounds by Hohenberg and Kohn (HK) [88], thus paving the way to the development of the density functional theory.

Their paper applies to the ground state of any electron gas experiencing external potential. Two theorems are proved in the paper:

1. The external potential $V(\mathbf{r})$ is (to within a constant) a unique functional of electron density $\rho(\mathbf{r})$. Since, in turn, $V(\mathbf{r})$ fixes the Hamiltonian therefore the full many-particle ground state is a unique functional of $\rho(\mathbf{r})$
2. The electron density $\rho(\mathbf{r})$ that provides minimum of the total energy functional is the true $\rho(\mathbf{r})$ of the corresponding solution of the Schrödinger equation.

The electronic energy in the HK formalism becomes

$$\langle \Psi | \, \hat{\boldsymbol{H}} \, | \Psi \rangle = E[\rho] = F[\rho(\mathbf{r})] + \int V(\mathbf{r})\rho(\mathbf{r}) d\mathbf{r} \tag{2.8}$$

where the functional $F[\rho]$ can further be split into Hartree energy $E^H[\rho]$, exchange and correlation $E^{XC}[\rho]$, and kinetic energy $T[\rho]$. Note that the latter two functionals are undefined in the original theory.

2.1.3.2 Kohn–Sham Equations

According to the Kohn-Sham (KS) formalism [104] of DFT the complicated system of N electrons can be substituted with N-systems of non-interacting electrons experiencing external effective potential V^{eff}. This greatly simplifies the problem as now one is only faced with solving the SE for one electron systems. Furthermore Kohn-Sham proposed that the unknown kinetic energy functional can be approximated by the kinetic energies of the KS orbitals. The difference between those kinetic energies is accommodated by the E^{XC} term which remains unknown and for which an approximation should be made.

The electron density is obtained from electronic orbitals

$$\rho(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2 \tag{2.9}$$

The KS total kinetic energy for a system is just a sum over all individual electron kinetic energies

$$T^{KS} = -\frac{\hbar}{2m_e} \sum_i \langle \phi_i | \nabla_i^2 | \phi_i \rangle \tag{2.10}$$

Hence the total energy of the system is

$$E^{KS}[\rho] = T^{KS}[\rho] + V^H[\rho] + V^{ext}[\rho] + V^{XC}[\rho] \tag{2.11}$$

where it is worth noting that non-interacting electronic orbitals have been reintroduced for the kinetic energy, while the forms of the second and third terms are known from electrostatics and the V^{XC} term contains the additional component of kinetic energy due to exchange and correlation which is discussed in section 2.1.3.3.

The ground state energy resulting from minimisation of Kohn-Sham equations is obtained using self consistent field method (SCF). The procedure begins using guess density, followed by the calculation of the effective potential. Kohn-Sham equations are then solved for every orbital resulting in a new density. The process is repeated until convergence in density is reached.

2.1.3.3 Exchange–correlation Functionals

Development of accurate exchange and correlation functionals is an active area of research. Here two well-established methods are presented, namely, Local Density Approximation (LDA) and Generalized Gradient Approximation (GGA) [14, 139].

LDA is the simplest approximation to the E^{XC} functional where the exchange and correlation are modelled as in the homogeneous electron gas model (HEG)

$$E^{XC}_{LDA} = \int \rho(\mathbf{r}) \epsilon^{XC}[\rho(\mathbf{r})] d\mathbf{r} \tag{2.12}$$

where $\epsilon^{XC}(\rho) = \epsilon^X(\rho) + \epsilon^C(\rho)$ is linearly combined exchange and correlation

energy per electron of a HEG. The exchange functional for Thomas-Fermi model has known analytical form due to Dirac [52]. The correlation term is usually obtained by fitting to accurate Monte Carlo calculations. The LDA performs surprisingly well given its simplicity mainly due to fortuitous cancellation of errors. LDA tends to overestimate ϵ^C and underestimate ϵ^X.

GGA improves upon LDA by including a dependence on the gradient of the charge density. The optimal functional form for GGA is still a matter of debate. There are two distinct approaches taken in the development of GGAs. Semi-empirical GGA's, which usually works well for small molecules but fails for bulk systems such as metals (e.g. BLYP [108, 125]). Alternatively, a numerical form is obtained from the first principles (e.g. PBE [138]).

It is worth noting that selection of exchange-correlation functional should be guided by the problem at hand. For example, adsorption on metal surfaces is best described by the BLYP functional even though it underestimates the atomization energies and overestimates the equilibrium volume [177]. On the other hand PBE yields sensible results for most metal properties and metals, but adsorption energy is grossly overestimated as compared with experiments.

2.1.3.4 Pseudopotentials

Electrons in the system can broadly be divided into core and valence electrons. Core electrons are those closer to nuclei, are tightly bound and are not involved in bonding. Valence electrons are the opposite and are critical to valid description of chemical bonds.

In principle, the KS orbitals can be computed for every electron in the system. However this leads to some difficulties as core electron wavefunctions oscillate rapidly due to high density gradient. The proper description of those wavefunctions would require highly accurate expansion in the basis set (see 2.1.3.5) which would render any scheme based on plane waves computationally inefficient for practical purposes.

Fortunately, one can exploit the fact that core electrons do not participate in bonding and can be combined with nuclei and represented as external potential known as the pseudopotential. Consequently, the KS orbitals can be solved just for valence electrons in the effective potential now including core electrons. The idea of pseudopotentials precedes DFT by over 30 years and is now one of the

most widely applied methods in computational physics [157].

The exact formulation of pseudopotential is an approximation. The pseudopotential is not only a screened Coulomb potential due to nuclear charges, but should incorporate relevant physics such as relativistic nature of core electrons. It should also properly model exchange-correlation interaction between core and valence electrons. Hence it is imperative that the pseudopotential itself depends on the electron density of all electrons. This requirement introduces non-local pseudopotentials which have the following form [182]

$$U^{ext}(\mathbf{r}) = \int U(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}')d\mathbf{r}' \tag{2.13}$$

2.1.3.5 Basis Set Expansion

The variational method requires that the single electron orbitals must be expanded using some basis set. The set must satisfy certain criteria such as smooth balance between expansion accuracy and computational efficiency.

The most common choice for the expansion of the single electron wavefunction in a periodic system, such as a crystal, is plane wave basis set. It follows naturally from Bloch's theorem which considers electrons in the periodic potential [28]. In the Bloch case the solution to the Schrödinger equation takes the form of a plane wave $e^{i\mathbf{kr}}$ modulated by a periodic function $u(\mathbf{r})$

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{kr}}u_{\mathbf{k}}(\mathbf{r}) \tag{2.14}$$

where $\mathbf{k}$ is a wave-vector of a plane wave and i is the imaginary unit.

Here, the periodic function is expanded in plane waves with wave-vector $\mathbf{G}$

$$u_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} C_{\mathbf{Gk}} e^{i\mathbf{Gr}} \tag{2.15}$$

where Ω is a unit cell volume.

when the above equations are combined, it provides a simple basis set for expansion

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} C_{\mathbf{Gk}} e^{i(\mathbf{G}+\mathbf{k})\mathbf{r}} \tag{2.16}$$

The summation in equation 2.16 is, in principle, over infinite number of plane

waves. For practical purposes, the summation is usually truncated at $\mathbf{G}_{max}$ since the coefficients of plane waves with small kinetic energies contribute the most to the ground state wavefunction.

The truncation introduces error into calculation of energy and forces. However the error can be assessed by performing convergence testing to find E_{cutoff} such that the total energy in converged within the required tolerance.

$$|\mathbf{G} + \mathbf{k}|^2 < \frac{1}{2}\mathbf{G}_{max}^2 = E_{cutoff} \tag{2.17}$$

While plane wave basis set is a logical choice for periodic condensed systems there are other possibilities which are often more appropriate for a problem at hand.

Linear combinations of atomic orbitals (LCAO) are suitable for constructing wave functions for molecules as the true wave function might be similar to these basis functions. LCAO basis functions are simply solutions from solving the Schrödinger equation for an isolated hydrogen atom (e.g. Slater type orbitals (STO) [168]). Although they are orthogonal when centred on one atom, they are not when centred on different nuclei. As a consequence it is difficult to systematically reduce error in total energy due to the usage of finite basis set - increase in number of basis functions does not guarantee higher accuracy.

Another popular choice are so called Gaussian type orbitals (GTO) proposed in 1950 by Boys [32]. Even though they are *less physical* than STO (no cusp at nuclei centre) they are better suited for modern computation thanks to the fact that the product of two GTOs is just another GTO. GTOs form a complete basis set even though they are not orthogonal. They realise the usual decomposition into radial part $R_l(\mathbf{r})$ and a spherical harmonic $Y_{lm}(\theta, \phi)$

$$\Phi(\mathbf{r}) = R_l(\mathbf{r})Y_{lm}(\theta, \phi) \tag{2.18}$$

There are many more different choices for basis sets however this is beyond the scope of this work.

2.1.3.6 Software implementing DFT

DFT is well established and de facto a standard for electronic structure calculations. The popular commercial DFT packages implemented with a plane

wave basis set used by the community are Vienna Ab initio Simulation Package (VASP) [105] and CASTEP [160]. The Quantum Espresso is another popular package which is freely available under GNU General Public License [71].

2.1.4 Tight-binding Methods

In this section a brief description of tight-binding (TB) method is provided for completeness. TB theory provides a bridge between quantum mechanical DFT and methods used in the development of classical interatomic potentials [182].

The theory is built on the idea that electrons are tightly bound to atoms and that the bonding process does not affect wavefunction in an appreciable manner. In general, those assumptions works reasonably well for strongly covalent systems but gradually break down as we move towards a free electron gas picture. Still, the TB method works with simple metals which are bonded with sp-electrons.

Similarly to DFT, TB employs variational method to find wavefunction expansion coefficients which minimise the energy of the system. The wavefunction is typically expanded in the LCAO basis set

$$\phi(\mathbf{r}) = \sum_{i,\alpha} c_\alpha^i \phi_\alpha^i(\mathbf{r}) \tag{2.19}$$

where α labels $s, p, d...$ orbitals, c_α^i are expansion coefficients and ϕ_α^i is hydrogen like orbital centred on atom i.

The assumption of tight binding justifies use of local LCAO basis set even though individual orbitals are not orthogonal when centred on different atom. It is assumed that the magnitude of resulting error is negligible.

As usual, the Hamiltonian matrix is constructed and the eigenvalue problem solved. Integrals resulting from the kinetic energy operator involve either one or two atoms and are called *one-centre* and *two-centre* respectively. To simplify the Hamiltonian matrix further, TB introduces a critical assumption that the total potential can be written as a sum of decoupled atom-centred potentials.

$$U(\mathbf{r}) = \sum_{i}^{N} U^i(\mathbf{r} - \mathbf{R}_i) \tag{2.20}$$

where i labels atom at position $\mathbf{R}_i$ and U^i is i-th atom centred potential. This is

rather speculative and cannot be really justified as the total energy is a function of the electron density in general. The assumption in equation 2.20 leads to simplified terms in the Hamiltonian matrix

$$U^{ij}_{\alpha\beta} = \sum_k \left\langle \phi^i_\alpha \right| U^k(\mathbf{r} - \mathbf{R}_k) \left| \phi^j_\beta \right\rangle \tag{2.21}$$

where i, j, k label atoms and α, β labels LCAO atom centred orbitals ϕ. The streamlined Hamiltonian matrix involves now *one-centre* ($i = j = k$), *two-centre* and *three-centre* ($i \neq j \neq k$) integrals.

The TB method assumes that due to tight-binding *one-centre* integrals can be ignored because they do not change from the atomic to the condensed phase, and that *three-centre* integrals are close to zero. The *two-centre* integrals are replaced with empirically fitted functions which depend on atomic coordinates only.

One must notice that dropping *three-centre* integrals is clearly invalid when k atom is close to i and j atoms as integral in eq. 2.21 is likely to be non-negligible hence limiting applicability of the TB method. The *two-centre* integrals can be parameterised using method proposed by Slater and Koster [167].

The resulting Hamiltonian matrix is sparse resulting in efficient inversion. In plain TB there is also no iteration required to converge electron density although there exist other TB flavours, such as density functional tight binding [55], where charge density is computed in a self-consistent manner . Another major drawback of TB models is their lack of transferability in comparison with DFT due to parameterised analytical functions used to approximate *two-centre* integrals. e.g. if the parameterisation is done by fitting to data, then the missing three-body terms will be approximated with two-body effects.

2.1.4.1 TB and classical interatomic potentials

As alluded earlier TB provides some theoretical background for the development of classical potentials. The discussion in this section follows closely [182].

The total energy can be computed from the density of states (DOS) by integrating over all available states. Since DOS contains all the information we need about the bonding in the material, therefore it can be used as a main building block during the development of the potential. Furthermore it has been shown that DOS can be approximated from its moments $\mu(n)$ where n labels n-th moment

[46].

The moments of the DOS can cheaply be computed from TB Hamiltonian matrix elements. The n-th moment is obtained from the Hamiltonian without diagonalisation

$$\mu_\alpha^i(n) = \left\langle \phi_\alpha^i \right| \hat{\boldsymbol{H}}^n \left| \phi_\alpha^i \right\rangle \tag{2.22}$$

where $\hat{\boldsymbol{H}}^n$ is raised to n-th power.

The zeroth moment is the normalisation of the DOS and the first moment is the total electronic energy of the system therefore they are not not particularly useful when one attempts to describe bonding. However higher order moments provide some insight into shape and size of the electronic density of states thus allowing to distinguish between different crystal structures. Consequently the approximate model can be build from second and higher order moments.

2.1.5 Coupled Cluster

Coupled Cluster theory is perhaps the most popular method to obtain high accuracy approximate solution to the time independent Schrödinger equation. Computationally it is very expensive and limited to systems where interaction is between small numbers of atoms. Still, it is very useful as can provide benchmark results for other methods such as DFT or high quality data which can be used to parameterise classical or machine learning interatomic potentials.

Coupled cluster (CC) method has been introduced in 1960 by Čižek and Paldus [40–42] using quantum field theory toolkit. The theory allows one to systematically improve the description of electronic exchange and correlation effects beyond the Hartree-Fock approximation. The balance between accuracy of the method and computational efficiency is obtained by truncating wavefunction ansatz given by

$$|\Psi_{CC}\rangle = e^{\hat{T}} |\Phi_{HF}\rangle \tag{2.23}$$

where Φ_{HF} is a Slater determinant composed of Hartree-Fock orbitals and $\hat{T}$ is the cluster operator defined as

$$\hat{T} = \sum_m^n \hat{T}_m \tag{2.24}$$

here n truncates the order of the approximation by controlling level of correlations

(e.g. $n = 2$ treats pair correlations exact within the given basis set). It is assumed that the lower level correlations are more important - pair correlations are more dominant than triplets and so on.

The cluster operator $\hat{T}$ is composed of creation and annihilation operators as well as corresponding (unknown) amplitudes. The action of $\hat{T}$ on the ground state wavefunction $|\Phi_{HF}\rangle$ results in a linear combination of excited determinants. In the popular coupled clusters singles and doubles theory (CCSD) the cluster operator is approximated as $\hat{T} = T_1 + T_2$. The triplet excitations are rarely treated explicitly, instead perturbative approach is used resulting in CCSD(T) theory. The latter theory is currently considered as a sweet spot between accuracy and the computation effort required [209].

2.2 Classical Interatomic Potentials

The interatomic potential describes interaction between atoms in the condensed phase. In principle, it should capture potential energy arising from the nucleus-nucleus interaction as well as both potential and kinetic energies from the ground state electrons. It is desirable that model does not only reproduce required properties of interest, such as cohesive energy or elastic constants, but is general enough to study other phenomena to which was not originally designed for. While classical interatomic potentials lack accuracy as compared with ab initio methods, such as density functional theory, they compete by striking the right balance between reliability and computational efficiency. Therefore they can be used to tackle large scale problems which are unachievable for other methods.

The general form for the potential function for N atoms with atomic coordinates $\{\mathbf{r}_i\}$ can be written as:

$$U_{tot} = U(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.25}$$

The justification to write such a function in terms of atomic coordinates only, and neglecting electrons, is provided in section 2.2.0.3. Here we note that a more workable form can be written down thanks to some physical invariants (see 2.2.0.4) and Cauchy's basic representation theorem [37]

$$U_{tot} = U(\{r_{ij}\}) \tag{2.26}$$

where $\{r_{ij}\}$ represents the set of all interatomic separations between particles in

the system. In other words the potential energy function is uniquely defined by the set of all labelled inter-particle distances. The discussion of this result can be found in [5, 37, 182, 193].

It is often the case that the total potential of the system U_{tot} is written as a series expansion of two-body, three-body, up to N-body interactions:

$$U_{tot} = \sum_i U_1(\mathbf{r_i}) + \sum_{i,j} U_2(\mathbf{r_i}, \mathbf{r_j}) + \sum_{i,j,k} U_3(\mathbf{r_i}, \mathbf{r_j}, \mathbf{r_k}) + ... \tag{2.27}$$

While the basic representation theorem legitimises this expansion it is often the case that it is truncated for practical reasons. The full expansion intuitively make sense but the truncation is only justified by the success of interatomic potentials when validated by the experimental data.

Another common form is obtained by partitioning of the total potential energy into individual atom contributions

$$U_{tot} = \sum_i U_i \tag{2.28}$$

This form is directly obtainable from eq. 2.27 by making use of the symmetry of the potential with respect to permutations of atomic positions. [182].

However, one must be careful when associating physical meaning to the energy of an individual atom given the cluster expansion is not unique. This is simply illustrated by noting that electrons have energy, and there is no unique way to assign electrons to atoms. Most importantly the total force on the atom is invariant of this partitioning thanks to the Newton's third law.

2.2.0.1 Forces

Without loss of generality, the force on atom i can be written as a sum of pairwise forces

$$\mathbf{F}_i = \sum_{\substack{j \\ j \neq i}} \mathbf{F}_{ij} \tag{2.29}$$

where $\mathbf{F}_{ij}$ represents contribution to the force on atom i due to atom j. Such a decomposition is true in general but not unique.

From the conservation of the linear and angular momentum of a system it can be

shown that the total force on the atom can be decomposed as a sum of pairwise forces and that those forces are symmetric in nature ($\mathbf{F}_{ij} = -\mathbf{F}_{ji}$) and are aligned along the relative position vector between interacting atoms [182]. These results are often referred as weak and strong law of action and reaction respectively. While the above seem obvious for the simple two-body potential it is worth noting that it still holds true for more complex models.

The force on atom i can be obtained from the total potential and together with eq. 2.26 one can show that the force obey weak and strong laws of reaction:

$$\mathbf{F}_i = -\frac{\partial U_{tot}}{\partial \mathbf{r}_i} = \sum_{\substack{j \\ j \neq i}} \frac{\partial U(\{r_{ij}\})}{\partial r_{ij}} \frac{\mathbf{r}_{ij}}{r_{ij}} \tag{2.30}$$

2.2.0.2 Atoms as classical particles

The thermal de Broglie wavelength gives us a way to estimate the temperature at which atoms behave as classical particles. If the de Broglie wavelength λ_{th} is much smaller than the interatomic distances one can assume that Maxwell-Boltzmann statistics is applicable. Below condition must be satisfied:

$$\lambda_{th} = \sqrt{\frac{2\pi\hbar^2}{m k_b T}} << \left(\frac{V}{N}\right)^{1/3} \tag{2.31}$$

where V is the volume of the system at temperature T and N is the number of particles with mass m.

Since most of the atoms behave classically at temperatures just above few Kelvins we can approximate forces, which are quantum mechanical by nature, using an interatomic potential energy function. However, one must be careful when designing models for light atoms, such as H or He, to be used at low temperatures.

2.2.0.3 Born-Oppenheimer Approximation revisited

The BOA has been introduced in section 2.1.0.1 as a way to decouple electronic and nuclear wavefunctions. Another extremely useful consequence of it is that it allows us to write the total potential field for the Schrödinger equation of the

nuclei without explicit dependence on the electrons [182]

$$U(\mathbf{r}) = U^{ZZ}(\mathbf{r}) + \epsilon_0(\mathbf{r}) \tag{2.32}$$

where $U^{ZZ}(\mathbf{r})$ represents usual nucleus-nucleus Coulomb interactions and $\epsilon_0(\mathbf{r})$ is an unknown potential field due to the electrons. Note that both functions depend on the nuclear coordinates.

In essence BOA provides firm physical background for the development of the interatomic potentials assuming that atoms can be treated as classical particles.

2.2.0.4 Invariance with respect to changes of reference frame.

The law of physics restricts the mathematical form of the interatomic potentials. In particular they must obey certain invariance laws.

Any system response to the deformation must be independent of the frame of reference. As a consequence, mathematically we require that any translation $\hat{\mathbf{T}}$ or rotation $\hat{\mathbf{R}}$ applied to the atomic coordinates leave function unchanged:

$$U(\hat{\mathbf{T}}\hat{\mathbf{R}}[\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N]) = U(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.33}$$

2.2.0.5 Invariance with respect to permutation.

Intuitively, since atoms of the same species are indistinguishable one requires that the potentials energy function is unchanged under permutation $\hat{\mathbf{P}}$ of the atomic coordinates.

$$U(\hat{\mathbf{P}}[\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N]) = U(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.34}$$

2.2.0.6 Invariance with respect to the inversion of space.

This requirements comes from the fact that all Hermitian Hamiltonians in quantum mechanics with the following form

$$\hat{\boldsymbol{H}} = \hat{T} + V(\{\mathbf{r}\}, \{\mathbf{R}\}) \tag{2.35}$$

have parity symmetry and all Hamiltonians describing interatomic bonding are Hermitian [22, 182]. Here $\{\mathbf{r}\}, \{\mathbf{R}\}$ label electronic and nuclear coordinates

respectively.

The invariance with respect to the inversion of space can mathematically be represented as

$$U = U(-\mathbf{r}_1, -\mathbf{r}_2, ..., -\mathbf{r}_N) = U(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.36}$$

2.2.0.7 Cutoff function

Most of the potentials presented in this section have an infinite range of interactions. While physically correct, in practice it is not desirable. First, it increases number of computations as pair interactions grow as N^2 where N is the number of particles in the system. Second, the simulated system is finite and often periodic boundary conditions are used. In the latter case the infinite range of interactions would lead to self-interaction between particles which is clearly unphysical.

The assumption of short-range interactions holds for many systems with a few exceptions such as ionic systems with long range Coulombic forces. In covalent systems the electrons are localised in bonds, metals have localised electron states in interstitials while van der Waals interactions are weak in nature due to shielding effect. Even though the Coulomb interactions (i.e., between protons and electrons) are always present in those system, the cancellation of opposite charges often results in mutual cancellation.

The common workaround is to introduce the cutoff function which truncates interactions at selected distance from the particle. This is often justified as long range interaction are often negligible. For the pair potentials it is simply a product of the original potential function and the cutoff function. The other possibility is to incorporate cutoff into the original potential. While it can be possibly more computationally efficient it limits the mathematical form because the potential function must go smoothly to zero at the cutoff distance.

It is worth noting that the simple truncation of the potential at the cutoff distance will introduce a jump in the potential. This may cause an unphysical behaviour in simulations and violate the energy conservation. The simple shift of the pair potential energy function can alleviate some of those issues. Still, the derivative of the potential function might not tend smoothly to zero at the cutoff distance. The selection of the cutoff distance is a balancing act where on one hand the

minimal cutoff is preferred due to the computational efficiency on the other hand it must be long enough to capture all the essential physics.

2.2.1 Two-body Potentials

Simple pairwise potentials can work remarkably well to describe systems where particles interact by the weak dispersion forces when other many body contributions to the total potential energy are negligible. They are often capable of capturing essential physics like melting, freezing, condensation or critical and triple points. Another advantage which comes from their simplicity is that they can often be fitted using experimental data.

$$U_{tot} = \sum_i \sum_{j>i} U_2(r_{ij}) \tag{2.37}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ and fig. 2.1 shows geometry of this interaction.

The commonly encountered limitations are the fact that they cannot distinguish between atoms in the perfect bulk and those with a defect close to them. Crystallographic defects disturb the regular crystal lattice and occur in all crystalline materials. For example, a vacancy defect is a missing ion from one of the lattice sites. This lack of the environmental dependence causes all pairwise interactions to be treated in the same way, while in real materials bonding strength will change due to the Pauli principle. The overlap of wavefunctions will promote additional electrons to occupy higher energy orbitals. In general, the bond gets weaker as local atomic environment is becoming crowded and stronger when atoms are removed from its closest vicinity. Another limitation is their inability to capture environmental dependence of bonding which is important, for example, when simulating transition metals or covalent systems. This limitations strongly affects properties such as vacancy formation and surfaces energies.

In covalent materials the directional bond is a consequence of overlapping atomic orbitals (i.e., electrons are shared within the bond). The shape of overlapping orbitals constraints possible spatial arrangements of atoms. In metals, such as those with body-centred cubic structure (bcc), the localisation of charges in the interstitial region results in formation of directional bonds due to the overlap of partially filled d-bands.

It is worth noting the difference between symmetry which arises from the

arrangement of atoms on a crystal lattice and directional bonding as discussed above.

A mathematical constraint of pair potentials is that they satisfy the so called *Cauchy relation*. The generalised Hooke's Law gives stress ($\boldsymbol{\sigma}$) strain ($\boldsymbol{\epsilon}$) relation for 3-dimensional systems

$$\sigma_{ij} = c_{ijkl}\epsilon_{kl}$$

where $\mathbf{c}$ is the *elastic tensor*. The symmetries of the stress and strain tensors allow to rewrite the 81 component elastic tensor in a compacted form (Voigt notation) giving rise to *elastic matrix* $\mathbf{C}$ with a maximum of 21 independent elastic constants C_{ij}. The number of elastic matrix components is further reduced due to the symmetries of a crystal.

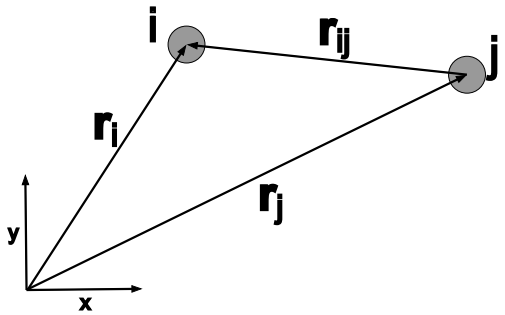


Figure 2.1: Two-body diagram.

For example materials with the cubic symmetry have only three independent elastic constants: C_{11}, C_{12} and C_{44}. For simple interatomic potentials it is possible to compute elastic constants directly using known formulas (eq. 11.135 in [182]). Specifically for pair potentials the Cauchy relation states that the ratio of C_{12} and C_{44} elastic constants is unity. Therefore the Cauchy relation is a direct artefact of a pair potential and does not hold in general for real materials.

The presence of many-body interactions in real materials results in a violation of the Cauchy relation and the magnitude of it gives insight on the nature of the bonding. This relation is, unsurprisingly, never satisfied for metals. However, it is often true for materials where either ionic bonding or van der Waals interactions are dominant.

Uncomplicated two-body interatomic potentials can be constructed using repulsive part only but in this case no molecular binding can occur. Nevertheless, even the simplest hard sphere potential shows first order phase transitions [90]. It is usually convenient to write down a pair potential as a sum of attractive and repulsive terms.

2.2.1.1 Lennard-Jones

Lennard-Jones (LJ) [98, 99] is arguably the most popular pair potential. It consists of attractive r^6 and repulsive r^{12} terms

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.38}$$

where r is the distance between atoms and σ and ϵ are fitting parameters. The attractive term comes from the dispersion forces due to fluctuating dipoles and describes interaction between two electronic clouds. The attractive term was obtained by Taylor expanding the potential energy of two interacting three-dimensional isotropic quantum harmonic oscillators in $1/r$ from the the second-order perturbation theory by Fritz London in 1930. The repulsive r^{-12} term on the other hand is chosen for computational efficiency rather than some deeper underlying theory and is meant to describe Pauli repulsion. The LJ has been shown to be too strong at short distances resulting in low compressibility and is often replaced with more accurate ZBL (2.2.1.4) potential. This is a general problem faced by many interatomic potentials were repulsive interaction is facilitated by the electron-electron repulsion while at close distances screened nucleus-nucleus interactions becomes dominant. Two fitting parameters σ and ϵ represent the approximate size of the particle and the depth of the potential well respectively. LJ potential describes reasonably well noble gasses such as argon. Noteworthy, there exist other LJ type potentials but with exponents different than 6 and 12. Those are referred in literature as Mie potentials. The LJ potential is often used as a building block for a more sophisticated force fields such as OPLS [100, 101] to describe more complex molecules. The LJ potential is most commonly used to study general class of effects, such as ductile failure in hexagonal closed packed (hcp) crystal structures, rather than properties of a particular material. For practical calculations, LJ is combined with some cutoff, and phase stability in Lennard-Jones systems is extremely sensitive to choice of cutoff [115].

2.2.1.2 Buckingham

The Buckingham potential [34] was designed to improve on the unphysical r^{-12} repulsive term in the original LJ potential. It does it by introducing an exponential term which is partially justified as a repulsion between two electronic

clouds [97]

$$V(r) = A\exp^{-Br} - \frac{C}{r^6} \tag{2.39}$$

where A, B and C are the fitting parameters and r is the distance between atoms.

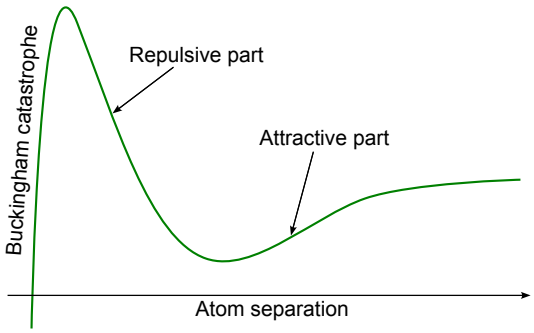


Figure 2.2: Buckingham catastrophe.

While it improves the repulsive part it suffers from the "Buckingham catastrophe". This is because the repulsive part converges to a constant for small distances while the attractive term diverges. This examples nicely illustrates that even for simplest of potentials some compromises have to be made. While the Buckingham potential is arguably more physical, it is not as popular as LJ. Historically this is mostly because the exponential term used to be much more expensive to compute. However, modern processors can compute exponentials much faster as the trigonometric functions are implemented on the hardware level.

2.2.1.3 Morse

Morse potential [130] can be used to describe the chemical bond of the isolated molecule or cases where the attractive interaction in the condensed system is composed of such bonds. The Morse potential provides better description for some of the properties of hcp and fcc metals compared to LJ potential. Yet it still cannot capture many-body effects which are essential for the proper description of metals.

The Morse potential improves on the quadratic potential of the quantum mechanical oscillator by including bond breaking and bond anharmonicity.

$$V(r) = D_e\big(1 - \exp^{-a(r-r_e)}\big)^2 \tag{2.40}$$

Here, D_e is a potential energy well depth, r_e is an equilibrium bond distance and a determine the width of the potential.

By expanding the squared term it can be shown that the Morse potential is

composed of attractive and repulsive terms in line with other pairwise potentials.

2.2.1.4 Universal Ziegler-Biersack-Littmark

Most interactions in condensed systems are facilitated by the interacting electron clouds. However, as nuclei are getting closer the Coulombic repulsion is becoming more important and can be modelled by the screened Coulombic potential of the following form

$$V(r) = \frac{1}{4\pi\epsilon_0}\frac{Z_1 Z_2 e^2}{r}\phi(r/a) \tag{2.41}$$

where Z_1 and Z_2 are the nuclear charges, r is the distance between two nuclei and e is the electron charge. One popular choice for the screening function ϕ and the parameter a is one proposed by Ziegler, Biersack, Littmark in 1985 [213]

$$a = \frac{0.8854 a_0}{Z_1^{0.23} + Z_2^{0.23}}$$

where a_0 is the Bohr atomic radius and the screening function is given by

$$\phi(x) = 0.18175e^{-3.19980x} + 0.50986e^{-0.94229x} + 0.28022e^{-0.40290x} + 0.02817e^{-0.20162x}$$

where $x = r/a$.

The ZBL potential is particularly well suited for the modelling of high energy atom collisions typical for radiation damage simulations. Since it models only core-core repulsion, it needs to be combined with some attractive terms.

2.2.1.5 Covalent Bond potentials

One can build a potential focusing on pairwise covalent bonds, each of which is described by a pair potentials. This has an unusual feature of long-range and short-range repulsion, with intermediate range attraction. A drawback for MD is that one has to keep track of which atoms are bonded.

An example model of covalent bonding in silicon [1] is composed of repulsive pairwise interaction between screened ionic cores and and attractive bonding

term modelled using radial part of p hydrogen-like orbital

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} Ae^{-\alpha r_{ij}} - \frac{1}{2}\sum_{i=1}^{N}\sum_{n=1}^{4} Br_{ik_n}e^{-\beta r_{ik_n}} \tag{2.42}$$

here the first term represents nuclear repulsion and A, α, B, β are fitting parameters. The second term represents attractive interaction which is a sum over four valence electrons involved in the bonding process.

The rational behind this model is that instead of considering angles in the perfect diamond structure, one can recognise that the bonding energy can be also represented with a pairwise model. This is because for any triplet of atoms two nearest neighbours of a given atom are second neighbours of one another [1].

2.2.2 Three-body Potentials

The physical motivation behind three-body potentials is to improve on the description of directional bonding as compared to simple pairwise models. This is achieved by introducing the three-body term which can be associated with the angle between bonds.

$$U_{tot} = \sum_{i}\sum_{j>i} U_2(r_{ij}) + \sum_{i}\sum_{j>i}\sum_{k>j} U_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) \tag{2.43}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$

2.2.2.1 Stillinger-Weber

This was originally developed to describe diamond lattice of Si in 1985 by introducing explicit angle dependent term on top of the usual LJ pairwise interaction [174]

$$U_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) = \lambda \exp\left[\frac{\gamma}{r_{ij} - r_c} + \frac{\gamma}{r_{ik} - r_c}\right]\left(cos(\theta_{ijk} - \beta)\right)^2 \tag{2.44}$$

where the i-th atom is the centre atom in the three-body interaction. Here the exponential term is a cutoff function which smoothly terminates the potentials at a distance r_c while γ controls its shape. The second term can be interpreted as a strength of penalty for structure to move away from the tetrahedral angle.

When $\beta = cos(109.47°)$ the diamond structure will be the ground state.

By simply fixing the angle between atom triplets SW potential provides good description for a diamond structure. While it works well for a diamond bulk it fails in nearly in any other context. This is because the diamond structure is stabilised by favouring 109.47° angles rather than fourfold bonding. For example, the surface energies are wrong, as surface structures are unphysical, and the coordination number of a liquid is too low.

2.2.2.2 Tersoff

Tersoff potential [185] takes a different approach to SW. Instead of fixing angles to match diamond structure it attempts to modify the strength of the bond between atoms i and j based on the coordination number, angles and bond lengths of the local atomic environment. For the sake of brevity the simplified form is provided

$$U_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) = f_c\Big[f_R(r_{ij}) + \gamma_{ijk} f_A(r_{ij})\Big] \tag{2.45}$$

where f_R and f_A are repulsive and attractive terms respectively smoothed out by the the cutoff function f_c. While it may appear as a two-body potential it is still three-body in nature and belongs to a class of bond order potentials. The γ_{ijk} bond angle term consists of a summation over all atomic triplets centred at atom i. For this reason the attractive term is modified based on the bonding environment, for example the bond becomes weaker if the local atomic environment becomes too crowded.

The Tersoff potential is arguably more flexible as compared to SW as it is able to describe at least some of the chemical reactions as well as strong covalent bonding and systems that bond in different geometries than diamond structure. In particular, it improves on the description of liquid and amorphous phases of silicon but still overestimates its melting temperature. The Tersoff potential is also commonly used to simulate other elements with diamond cubic structure such as carbon and germanium.

2.2.3 Four- and Five-body Potentials

In general, as the number of atoms under consideration increase so the complexity of the model. One can quickly lose ability to appreciate what physical insight went into the development of the model. Moreover, it becomes much more difficult to determine whether the success of the model is due to the essential physics of the problem being captured or rather thanks to superior parametrisation procedure. Arguably, analytical four-body potentials demarcate this line as it becomes increasingly more difficult to parameterise higher order analytical potentials [182].

Arguably, the simplest five-body potential developed for silicon is presented below [1]

$$U_4 = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N} Ae^{-\alpha r_{ij}} - \frac{1}{2}\sum_{i}^{N}\sum_{k}^{4} Br_{ik_n}e^{-\beta r_{ik_n}} \tag{2.46}$$

where i and j sums are over all atoms and k sum is over all valence electrons of a bond energy. Given that silicon is tetravalent the summation is over four valence electrons which are shared within the bond resulting in a 5-body potential. A, B, α, β are fitting parameters. The first term represents pairwise repulsion between the nuclei (and the core electrons) separated by r_{ij}. The second term is a cohesion due to the valence electrons. r_{ik} is the distance between two ions on which the electron's orbitals are centred. Here one electron per bond is assumed for simplicity however this can be generalised to account for double bonds as well.

The model reproduces lattice parameters and relative energy differences between various silicon structures. It also describes well point defects and surfaces. Remarkably for such a simple model it outperforms, in those areas, other more complicated models [1]. Due to its mathematical formulation it fails to reproduce elastic shear constants. While the functional form of the potential aids appreciation of the underlying physics, the need to identify four neighbours somewhat complicates its implementation into MD software in an efficient manner when bond-breaking is allowed.

2.2.4 Many-body Potentials

Generally speaking any potential which incorporates higher than two-body interactions can be considered as a many-body potential. The three-body

potential has been discussed in the previous section. The expansion of the total energy up to the quadruplet term has been proposed in 1988 by Moriarty based on generalised pseudopotential theory [128, 129]. Obtaining higher order parametrisations is however challenging as the physical insight into the necessary functional form becomes blurred. In addition the truncation of the cluster expansion (equation 2.27) at lower order terms is only successful in cases where the convergence is fast. The alternative approach presented in this section is based on the idea that the energy of an atom depends on the local electron density to which all neighbouring atoms contribute. From now on the "many-body" term will refer to them.

The concept of the local density allows to modify the strength of the interatomic bond based on the local atomic environment. This is particularly useful for the description of surfaces and defects where large density changes occur.

The many-body potentials can be split into two groups as proposed by Carlsson in [35]. Pair functionals build density pairwise resulting in a spherical charge density approximation while cluster functionals will contain higher order terms. The many body potentials are usually based on tight binding models or local atomic embedding functions.

2.2.4.1 Pair functionals

Several different models have been developed in 1980s to tackle shortcomings of pair potentials. Arguably, two most successful are Finnis-Sinclair (FS) [64] model developed for transition metals based on tight binding theory and embedded atom method (EAM) [48] which is obtained from the density functional theory. There are a number of other pair functionals such as effective medium theory or glue potentials but most importantly they all share the same functional form:

$$U_{tot} = \sum_{i,j>i} \phi^{ij}(r_{ij}) + \sum_{i} F^{i}(\rho_i) \tag{2.47}$$

where ϕ^{ij} is a pairwise potential and F_i in an energy function which depends on density ρ_i - the local atomic environment of atom i. For pair functionals the local density term is approximated by

$$\rho_i = \sum_{j \neq y} f^{j}(r_{ij}) \tag{2.48}$$

where f^j is simply a pairwise function which effectively is a weighted count of neighbours.

The physical interpretation of equation 2.47 varies depending on the underlying theory. In the Finnis-Sinclair model the pairwise function ϕ^{ij} represents screened Coulombic repulsion between two nuclei each combined with its non-valence electrons. The functional F^i is given by

$$F^i = -A\sqrt{\rho_i} \tag{2.49}$$

where A is a positive fitting parameter. The attractive F^i term is obtained from the second-moment approximation to tight binding theory. The assumption is that the bond energy, in transition metals, scales with the width of the local density due to the filling of the d-bands [4, 64]. The width of any distribution is obtained from the square root of its second-moment. In this case the variance is approximated by the the linear superposition of pairwise interactions f^{ij} which represents sum of squares of electron hopping integrals between interacting atoms.

The functional F^i in the EAM model can be associated with the energy required to embed atom i in the homogeneous electron gas of density ρ_i. One can think of it as assembling a bulk atom by atom. Here, the local density term for each atom is approximated as a spherically symmetric cloud. The embedding function in the EAM model does not have a well defined functional form. In fact it is often a numerical function either obtained theoretically or experimentally.

Pair functionals provide vast improvement over simple pair potentials in the description of metals. Their simple, physically motivated, form is efficient to compute and only a factor of two slower as compared with pair potentials. They provide a good description for close-packed systems with full or nearly full d-bands.

2.2.4.2 Cluster functionals

The cluster functionals are built on the quantum mechanical idea of the bond order originally devised by Pauling. Simply put the bond order represents the strength of the bond and is related to the number of electron pairs between two interacting atoms. This number is strongly dependent on the local atomic environment.

Apart from the simplest s-orbitals, the electron density is a function of both distance and direction from the nucleus. Pair functionals assume spherically symmetric density which limits their applications. On the other hand cluster functionals combine atoms in groups of at least three thus allowing for non-spherical density function.

The most popular methods are based on analytical bond-order potentials (BOP) [91]. In a tight binding theory calculation of the bond order requires expensive diagonalisation of the Hamiltonian matrix. Analytical BOP methods aim to evaluate bond order from the moments of the electronic density of states. Even though moments can be obtained directly from the Hamiltonian, still from the practical perspective, the analytical BOP methods are too expensive and are further simplified in rather ad hoc ways.

Note that the Tersoff model, which was discussed in section 2.2.2.2 for convenience, can also be classified as BOP model. Another very popular approximation to BOP is a modified embedded atom method (MEAM) [12, 13]. As name suggests it is modified EAM with the aim to account for the directional nature of the electron density. The functional form remains the same as in any pair functional (eq. 2.47) but the density term includes corrections based on spherical harmonics to incorporate angular-dependent interactions. In this way description of covalent bonding is greatly improved in materials like silicon and also bcc metals with partially filled d-orbitals.

2.3 Molecular Dynamics

In the molecular dynamics (MD) simulation, atoms are treated as classical particles. The Newtonian equations of motion are integrated to obtain trajectory for each nuclei in the system

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i \tag{2.50}$$

where m_i represents mass of atom i and $\ddot{\mathbf{r}}_i$ is its acceleration. The external force acting on atom i is a combination of forces due to neighbouring atoms and optional external field.

The interaction between atoms is governed by the potential energy surface $V = V^{int} + V^{ext}$. Where V^{int} represents energy due to the interactions between atomic nuclei, and V^{ext} in an external potential field. The resulting force on atom i is

obtained from

$$\mathbf{F}_i = \frac{\partial V}{\partial \mathbf{r}_i} \tag{2.51}$$

The motion of atoms is obtained by integrating eq. 2.50. Various methods exists to numerically integrate Newton equation of motion however the most commonly used is the velocity Verlet algorithm (VV) [180].

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \tag{2.52}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{a}(t) + \mathbf{a}(t + \Delta t)}{2}\Delta t \tag{2.53}$$

The VV algorithm preserves all the important features of eq. 2.50, namely the conservation of the total energy of the closed system (or more precisely conservation of the approximation to the Hamiltonian which becomes exact in the limit $\Delta t \to 0$) and time-reversibility. The algorithm is also a symplectic integrator thus it provides excellent stability and conservation of energy [182].

The temperature in MD simulation is obtained from the time average of the kinetic energy

$$T = \frac{2}{3Nk_B}\overline{T} = \frac{2}{3Nk_B}\overline{\sum_{i=1}^{N}\frac{1}{2}m_i v_i^2} \tag{2.54}$$

It is worth emphasising that the lack of electron-phonon coupling in MD simulations is a source of potential error when heat transfer to and from electrons is essential to describe physical phenomena. This is the Born-Oppenheimer Approximation again.

So far the discussion in this section and the presented algorithms are applicable to the microcanonical ensemble, also known as NVE - where number of particles (N), volume (V) and total energy (E) are all constants (and total momentum is conserved).

To allow somewhat more direct simulations of experiments NVT and NPT ensembles can be used, where V represents simulation box volume, P its pressure and T temperature. Both require external thermostat to allow exchange of energy between the simulation box and the surroundings. The thermostat modifies eq. 2.50 and is required to sample true canonical ensemble. Two popular choices are Langevin and Nosé–Hoover thermostats. In general interatomic forces are functions of atomic positions, but when a thermostat is used then force becomes

a function of velocities as well.

The Langevin thermostat introduces random forces to the equation of motion thus mimicking energy exchange between the ensemble and the heat bath

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i - \gamma_i m_i \mathbf{v}_i + \mathbf{G}^i(t) \tag{2.55}$$

where γ_i is a damping constant and $\mathbf{G}^i(t)$ introduces random force on each atom i. The force coming from $\mathbf{G}^i(t)$ can be interpreted as random collisions of atom with some external medium of the heat bath. It is worth noting that due to its stochastic nature the Langevin thermostat is not deterministic: energy flows into the system from $\mathbf{G}^i(t)$ and is removed by γ_i. In consequence, moving backwards in time from a final state will not result in the initial microstate.

The Nosé–Hoover [89, 133] thermostat introduces a fictitious particle of mass M and momentum P to the simulated system. The equation of motion is modified to couple it with other atoms. The Nosé–Hoover thermostat is defined with the following equations

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i - \gamma m_i \mathbf{v}_i \tag{2.56}$$

$$\dot{\gamma} = \frac{1}{M}\left(\sum_i^N \frac{\mathbf{p}_i \cdot \mathbf{p}_i}{m_i} - 3Nk_BT \right) \tag{2.57}$$

where $\gamma = P/M$ is a damping coefficient which evolves according to the second equation above.

Perhaps more controversially, the thermostat can be used to alleviate accumulation of numerical errors during the simulation run. In even more extreme cases it can be utilised to stabilise simulations where the interatomic potential is not energy conserving (plain wrong in principle but surprisingly common) - of course any statistics obtained from such a run are simply invalid. For example, usage of Lennard-Jones potential with too short interaction cutoff and without appropriate truncation and shifting may result in unstable simulation.

In the same spirit as for the thermostat, the simulation box can also be coupled to a pressure bath with a Parrinello–Rahman (PR) barostat [137]. The PR barostat can be used along with a thermostat to simulate system in NPT ensemble which is the closest MD can get to the experiment. The equation of motion is further modified to include effects of changing the volume and shape of the simulation box.

2.3.0.1 MD potentials

All molecular dynamics simulations in this work are performed with the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package [189]. LAMMPS allows users to choose appropriate interatomic potential for the simulations at hand. Some simple potentials, such as Lennard-Jones require provision of just two numbers (σ and ϵ, see 2.2.1.1). The more sophisticated potentials (such as EAM 2.2.4.1) usually require a separate file which contains IP details. Those files can either be obtained from interatomic potentials repositories such as NIST [15, 80] or directly requested from authors.

For machine learning interatomic potentials 2.5 the situation is more complex given the field is still in the early days. In general, MLIP requires LAMMPS to be compiled by the user with appropriate MLIP plugin which is provided by the respective authors. Some popular and widely used MLIP packages such as GAP [11] are distributed with with LAMMPS source code. Once LAMMPS is compiled with the appropriate plugin it will allow the user to invoke MLIPs which are usually distributed as text files similarly to EAM.

2.4 Machine Learning

2.4.1 Overview

Machine learning (ML) can loosely be defined as a set of automated methods that can uncover patterns in data. Once ML model is trained it should be able to predict future data or aid decision making under uncertainty [131]. ML models are usually classified as either supervised, unsupervised or reinforcement learning approach. This work is mostly concerned with the supervised learning while the brief description of other two is given below for completeness.

In the *unsupervised* learning approach only inputs $\{\mathbf{x}_i\}_{i=1}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ are given without associated labels. Each training data point $\mathbf{x}$ is represented as a vector of numbers (e.g. width, length, depth and temperature of lochs is Scotland). The problem at hand must often be first translated into this numerical array before ML algorithms are employed. The purpose here is to uncover patterns in the data which might aid understanding of the problem. Typical examples are principal component analysis and clustering (grouping) of

data points.

The *reinforcement* learning is concerned with maximising intelligent agent's benefit given occasional reward or punishment signals (e.g. self-driving cars). Here the learning agent interacts with its environment by choosing an action at every discrete time step. The action taken affects the environment and in consequence affects the acting agent. If the action taken was beneficial to the agent it gets rewarded otherwise it gets punished. Through trial and error the model is refined to maximise model performance goals.

The goal of ML *supervised* method is to find a map which associates n training data observations $\{\mathbf{x}_i\}_{i=1}^n$ to a set of n labels $\{\mathbf{y}_i\}_{i=1}^n$. While some problems are inherently predisposed for ML approach (e.g. estimate body volume from people's height and weight), others might require more complex preprocessing. Furthermore, supervised methods are divided as classification (when $\mathbf{y}$ is categorical) or regression ($\mathbf{y}$ is real-valued).

ML models can be also be distinguished as *parametric* and *non-parametric models*. The former have a fixed number of parameters, while in the latter the number of parameters grow with the amount of data. Both types of models can suffer from, what is known in the literature as, the *curse of dimensionality*. In general, as the dimensionality of the problem increases so does the volume space of data. However, the increase in volume space is exponential often leading to sparse training data. One way to tackle this problem is to make some assumptions about the nature of the data distribution and incorporate them into a parametric model or a non-parametric model. It is worth noting that these classifications get blurred when using sparse kernel methods or neural networks with an increasing number of neurons.

2.4.2 Model Performance

The supervised method measures performance of the model using predefined rule. The common choices for regression models are mean absolute error (MAE as defined in eq. 2.58) and root mean square error (RMSE, eq. 2.59). For categorical data sets an error rate is used which gives proportion of cases where prediction is wrong.

$$MAE = \frac{1}{n}\sum_{i}^{n}|y_i - \hat{y}_i| \tag{2.58}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2} \tag{2.59}$$

where y and $\hat{y}$ are observed and predicted values respectively.

It is worth noting that MAE and RMSE are scale dependent therefore they should not be used to compare models trained using different training data sets. Out of both measures, the RMSE is more sensitive in picking up rare but large differences between true and predicted value. The difference between MAE and RMSE can provide insight on how good the model is on average (MAE) and how it copes with occasional outlying cases (RMSE).

2.4.3 Training Data

Good practice during the development of ML model is to split available data into *training*, *validation*, and *test* sets. In general the ML model consists of regression *weights* and model-specific *hyper parameters* (HP). The training set is used during initial fit with a supervised method of choice to find a set of weights given model HPs. Next, the model is tested on the validation set to further fine-tune model HPs. This two-step process can be used not only to optimise hyper parameters but also to prevent over-fitting to the training data set for example by cross-validation [131]. In the final step of the training process the test set is used to obtain unbiased evaluation of the performance of the model. Ideally, the test set is never used during optimisation of weights and hyper parameters (in this case it is often referred as hold-out data set).

2.4.4 Algorithms

2.4.4.1 Linear Regression

In this section the well known *linear regression* (LR) will be used to illustrate basic idea of machine learning supervised regression. The key properties of LR model are linearity in the parameter space $\{w_i\}_{i=1}^{N}$ as well as in the input space $\{x_i\}_{i=1}^{N}$. The LR model is defined as

$$y(\mathbf{w}, \mathbf{x}, \epsilon) = \mathbf{w}^T\mathbf{x} + \epsilon = \sum_{i=1}^{N} w_i x_i + \epsilon \tag{2.60}$$

where ϵ is the *residual error* between true and predicted values sometimes called *bias*. Here we assume that ϵ is normally distributed.

To optimise a model one has to choose a performance measure metric between predicted $\hat{y}$ and true values y from the training data set. The common choice used with linear models is the mean square error (MSE)

$$MSE = \frac{1}{n}\sum_{i}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{2.61}$$

Assuming MSE (eq. 2.61) as a performance measure the optimal set of weights is obtained by solving eq. 2.62 known as normal equation in the literature [26, 131]

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.62}$$

where $\mathbf{X}$ is called a *design matrix* (DM) that combines all inputs $\{\mathbf{x}_i\}_{i=1}^{N}$.

The closed form solution obtained from the normal equation approach has obvious advantages but also its limitations. Namely, while it provides the true minimum for LR for MSE performance metric it becomes intractible for problems with large number of parameters and training data points due to the size of the DM. An alternative approach for solving linear regression problem involves using online optimiser such as stochastic gradient descent algorithm.

The simplicity of the linear regression allows one to obtain some basic insight into a trained model. The relative magnitude of parameters indicate importance of input vector features. The sign of w_i reveals how the predicted value changes with respect to the change in the input feature.

The simple LR model can be extended to model nonlinear data sets by introducing basis functions ϕ, thus equation 2.60 becomes

$$y(\mathbf{w}, \mathbf{x}, \epsilon) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + \epsilon = \sum_{i=1}^{M} w_i\phi_i(\mathbf{x}) + \epsilon \tag{2.63}$$

There are many possible choices for basis function, some of them are discussed in [26]. The simplest basis function is an identity function $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$ in which case the models reduces to 2.60. The model extended with non-linear basis functions is still linear in the parameter space but nonlinear in the input space as every input vector $\mathbf{x}$ becomes $\{\phi_i(\mathbf{x})\}_{i=1}^{M}$.

Assuming that ϵ is a zero mean Gaussian random variable with variance σ_ϵ^2 and all training data are independent then the linear model can be motivated as the maximum likelihood solution. In this case the Gaussian conditional distribution is given by

$$p(t|\mathbf{w}, \mathbf{x}, \beta) = \mathcal{N}\big(t|y(\mathbf{w}, \mathbf{x}), \sigma_\epsilon^2\big) \tag{2.64}$$

and the corresponding likelihood function is

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\big(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \sigma_\epsilon^2\big) \tag{2.65}$$

where $\mathbf{X}$ is a data set of inputs and $\mathbf{t}$ collects all target variables. By maximising the log likelihood function one can obtain optimal weights and variance. This solution is equivalent to the one obtained from least squares.

The complexity of the model can be thus controlled by selecting appropriate number of basis functions in relation to the data set size. Alternative, and perhaps more flexible approach, is to introduce regularisation term which effectively controls model complexity. Then the total error function for least squares (2.61) now contains regularisation term λ

$$\frac{1}{n}\sum_{i}^{n}\big(y_i - \hat{y}_i\big)^2 + \lambda\mathbf{w}^T\mathbf{w} \tag{2.66}$$

For non-linear kernels the last term can be replaced with $\lambda\mathbf{w}^T\mathbf{K}\mathbf{w}$ where matrix $\mathbf{K}$ is a kernel matrix. The regularised solution to least squares problem is now given by

$$\mathbf{w} = \big(\lambda\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\big)^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{2.67}$$

where $\boldsymbol{\Phi}$ is a design matrix of post processed data with a basis function of choice. The regularisation parameter λ forces parameter values to shrink unless they are supported by the data.

2.4.4.2 Bayesian Linear Regression

In cases where the model is a linear combination of basis functions, the increase in the number of basis functions positively correlates with an increase in the maximum likelihood. This property leads to the phenomenon of overfitting - the model reproduces training data sets very well but fails to predict future

data reliably. As elucidated in section 2.4.4.1, an alternative method is to introduce regularisation parameter λ. In practice this shifts the problem from finding optimal set of basis functions to determining optimal value of λ. One possible solution is to split available data into training and validation sets. While this solution works well in cases where data are abundant it is wasteful when generating data is expensive. Alternative method which is common among machine learning practitioners is to use cross validation procedure to establish the value of λ [131]. While it works well it is not fully automated as it requires a user to provide appropriate range for the regularisation parameter.

The Bayesian approach to linear regression allows one to determine the regularisation parameter from the training data alone and it is fully automated [26]. The conjugate prior for the likelihood function (eq. 2.65) is given by the Gaussian distribution with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$

$$p(\mathbf{w}) = \mathcal{N}\big(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\big) \tag{2.68}$$

The posterior distribution of weights $\mathbf{w}$ given training data $\mathbf{t}$ is given by the Bayes' theorem

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\big(\mathbf{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\big) \tag{2.69}$$

where

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\Big(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_\epsilon^2}\boldsymbol{\Phi}^T\mathbf{t}\Big) \tag{2.70}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_\epsilon^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{2.71}$$

Since the posterior distribution is Gaussian, the vector of weights is given by $\boldsymbol{\mu}_N$. Furthermore by choosing zero centred Gaussian with variance σ_p^2 for the prior distribution the posterior distribution over the vector of weights is given by

$$\boldsymbol{\mu}_N = \frac{1}{\sigma_\epsilon^2}\boldsymbol{\Sigma}_N\boldsymbol{\Phi}^T\mathbf{t} \tag{2.72}$$

where

$$\boldsymbol{\Sigma}_N^{-1} = \frac{1}{\sigma_p^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{2.73}$$

Note that here the maximisation of the log likelihood function is identical to the regularised least squares solution (eq. 2.66) with $\lambda = {}^{\sigma_\epsilon^2}/_{\sigma_p^2}$.

The probabilistic approach provides valuable insight into the distribution of weights (given by eq. 2.69) which can be used as a measure of model uncertainty

about its weights. Moreover the predictive distribution is given by

$$p(t|\mathbf{x}, \mathbf{t}, \sigma_\epsilon, \sigma_p) = \mathcal{N}\big(t|\boldsymbol{\mu}_N \phi(\mathbf{x}), \sigma_N^2(\mathbf{x})\big) \tag{2.74}$$

where the predictive distribution variance takes the following form

$$\sigma_N^2(\mathbf{x}) = \sigma_\epsilon^2 + \phi(\mathbf{x})^T \boldsymbol{\Sigma}_N \phi(\mathbf{x}) \tag{2.75}$$

Thus each prediction has associated uncertainty with it. In other words, the model can tell us when it is unsure about its own predictions. This is a formidable feature as it could prevent deployment of the model on the data which are beyond model's capabilities.

So far the discussion above assumed that the noise in the data σ_ϵ^2 and the width of the prior distribution σ_p^2 are known. However this is not the case. Fortunately in the Bayesian treatment above parameters can be estimated in the iterative manner using framework known as the *evidence approximation*[1]

The predictive distribution in a fully Bayesian treatment is as follows [26]

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \sigma_\epsilon^2) p(\mathbf{w}|\mathbf{t}, \sigma_p^2, \sigma_\epsilon^2) p(\sigma_p^2, \sigma_\epsilon^2|\mathbf{t}) \mathrm{d}\mathbf{w} \mathrm{d}\sigma_p^2 \mathrm{d}\sigma_\epsilon^2 \tag{2.76}$$

here posterior distribution over σ_p^2 and σ_ϵ^2 is introduced and the integration is over both weights as well as above hyper parameters. Unfortunately there is no analytical solution to this problem. However if one assumes that the posterior distribution $p(\sigma_p^2, \sigma_\epsilon^2|\mathbf{t})$ is sharply peaked around values $\hat{\sigma}_p^2$ and $\hat{\sigma}_\epsilon^2$ than the the predictive posterior is obtained by integrating over weights only

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\sigma}_p^2, \hat{\sigma}_\epsilon^2) = \int p(t|\mathbf{w}, \hat{\sigma}_\epsilon^2) p(\mathbf{w}|\mathbf{t}, \hat{\sigma}_p^2, \hat{\sigma}_\epsilon^2) \mathrm{d}\mathbf{w} \tag{2.77}$$

The posterior distribution for σ_ϵ^2 and σ_p^2 is obtained from the Bayes' theorem

$$p(\sigma_\epsilon^2, \sigma_p^2|\mathbf{t}) \propto p(\mathbf{t}|\sigma_\epsilon^2, \sigma_p^2) p(\sigma_\epsilon^2, \sigma_p^2) \tag{2.78}$$

Assuming relatively flat prior the values of $\hat{\sigma}_p^2$ and $\hat{\sigma}_\epsilon^2$ are found by maximising the marginal likelihood function $p(\mathbf{t}|\sigma_\epsilon^2, \sigma_p^2)$ also known as the evidence function.

[1] In the statistic literature it is known as 'empirical Bayes' or 'type 2 maximum likelihood' or 'generalised maximum likelihood'. 'Evidence approximation' term is used in machine learning literature.

There are two approaches to finding the maximum of this function. The first option is to use an algorithm known as the expectation maximisation [26]. The second way is to analytically evaluate the evidence function and then set derivative wrt to hyper parameters to zero. Both solutions converge to the same result. *Ta-dah!* is using the latter.

2.4.4.3 Kernel Ridge Regression

The least squares problem can be entirely formulated in terms of the *kernel matrix* $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ (where $\mathbf{X}$ is a design matrix as defined in section 2.4.4.1) [26]. Here, the kernel matrix contains the inner products of all vector pairs. The matrix $\mathbf{K}$ can also be constructed for the non linear regression. In this case, its components are $\mathrm{K}_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. Therefore, the kernel function can be defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \tag{2.79}$$

The prediction is then given by

$$y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T\big(\mathbf{K} + \lambda\mathbf{I}\big)^{-1}\mathbf{t} \tag{2.80}$$

where the vector $\mathbf{k}$ is obtained using eq. 2.79 and $\mathbf{t}$ is a vector of training data.

Kernel ridge regression (KRR), in other words, is a regularised least squares where the inner product is replaced with the kernel function. This is an example of so-called *kernel trick*. It allows to reformulate algorithms in which the input vector enters only in the form of the inner product with the kernel of choice.

Our implementation of the KRR uses relatively little known tool called the Empirical Kernel Map (EKM) [154]. The EKM transforms vectors from the input space to the finite dimensional vectors which represent points in the kernel feature space. The EKM is defined as follows

$$\Theta_m(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{x}_n) \end{bmatrix} \tag{2.81}$$

Note that the EKM matrix $\mathbf{\Theta}$ differs as compared with the $\mathbf{K}$ matrix as $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \neq \Theta(\mathbf{x}_i)^T\Theta(\mathbf{x}_j)$. In principle, it is possible to restore the kernel

values since $\Theta'(\mathbf{x}) = \mathbf{K}^{-1/2}\Theta(\mathbf{x})$ however it defeats the purpose.

The EKM allows us to *kernalize* any algorithm which works on vectors by preprocessing the input vectors. In our case it permits usage of algorithms developed for the BLR with the KRR without any changes. Moreover, it also allows us to run our algorithms on the large data sets which would normally be impossible due to the high memory and computational efforts associated with the traditional approach of constructing $\mathbf{K}$ matrix. In the latter case the EKM is used with the linearly independent subset of the vectors living in the kernel space.

EKM constructs the covariance matrix such that its eigen structure is exactly that of the kernel matrix. Therefore it is possible, for example, to perform kernel principal component analysis (KPCC) and obtain the same results as with the full $\mathbf{K}$ matrix [121]. In the case when a large data set is present the approximate KPCC is possible since the reduced size $\boldsymbol{\Theta}$ provides an explicit mapping between input and feature spaces.

2.4.4.4 Generalised Linear Model

The generalised linear model (GLM) transforms the function linear in the parameter space $\mathbf{w}$ using a non-linear activation function f where its inverse f^{-1} is often referred to as a link function. Therefore, the GLM allows the model to be linked to the response variable by f^{-1} hence allowing the variance of each measurement to be a function of its predicted value. The GLM is simply defined as

$$y(\mathbf{x}) = f\left(\mathbf{w}^T\mathbf{x}\right) \tag{2.82}$$

In the remaining part of this section it is shown that the combination of the linear kernel and the quadratic kernel is equal to the 2nd order polynomial expansion of the scalar response variable (in our case it is an energy functional) in terms of the feature vectors as the independent variables. In this case, The GLM activation function is $f(x) = x + x^2$ and the resulting model is given by

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + \mathbf{x}^T\mathbf{w}\mathbf{w}^T\mathbf{x} \\ &= \mathbf{w}^T\mathbf{x} + \mathbf{x}^T\mathbf{W}\mathbf{x} \end{aligned} \tag{2.83}$$

The simplest possible kernel is a linear kernel also known as an inner or an identity

kernel. It is defined as

$$k_L(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' \tag{2.84}$$

When used with a model which is linear in the parameter space, k_L allows to cut computations by exploiting the following relation

$$y(\mathbf{x}) = \sum_b w_b \mathbf{x}^T \mathbf{x}_b = \sum_b \sum_\alpha w_b \mathbf{x}^{(\alpha)} \mathbf{x}_b^{(\alpha)} = \sum_\alpha \sum_b w_b \mathbf{x}^{(\alpha)} \mathbf{x}_b^{(\alpha)} = \sum_\alpha w_\alpha \mathbf{x}^{(\alpha)} \tag{2.85}$$

where index b goes over all basis vectors and α over vector components and w_b and w_α are model weights. It is straightforward to build complete orthonomal basis for the linear kernel as the number of basis vectors is simply equal to the dimension of the feature vector. One can see, that using a linear kernel simply reduces to regular regression and is equal to the first term in the second order polynomial GLM (eq. 2.83)

To obtain the second term in eq. 2.83 the quadratic kernel is used. The quadratic homogeneous kernel belongs to a more general group of polynomial kernels defined as $(\mathbf{x}^T \mathbf{x}' + c)^n$ with order n=2 and c=0

$$k_Q(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^2 \tag{2.86}$$

Note that, the higher order polynomial kernels may improve data fit but too often leads to overfitting especially for small data sets. Henceforth we restrict our use to n=2 which equals the order of the polynomial GLM expansion.

Contrary to the case of the linear kernel, the construction of the basis vectors $\{\mathbf{x}_b\}$ for the quadratic kernel is non trivial. One way is to use a modified on-line sparsification algorithm to select nearly linearly independent basis from the set of all available descriptors [56]. In our implementation the original algorithm proposes candidate basis vector, next the covariance matrix is constructed and tested for positive definiteness. If this test fails the candidate is rejected, otherwise it is added to the basis.

Even after the sparsification process the constructed basis is often too large to be efficiently used for predictions. Here it is presented a simple extension of the

linear kernel trick (eq. (2.85)) to the quadratic kernel

$$\begin{aligned} y(\mathbf{x}) &= \sum_b w_b (\mathbf{x}^T \mathbf{x}_b)^2 \\ &= \sum_b w_b \Big(\sum_\alpha \mathbf{x}^{(\alpha)} \mathbf{x}_b^{(\alpha)} \Big)^2 \\ &= \sum_b w_b \sum_\alpha \mathbf{x}^{(\alpha)} \mathbf{x}_b^{(\alpha)} \sum_{\alpha'} \mathbf{x}^{(\alpha')} \mathbf{x}_b^{(\alpha')} \\ &= \sum_\alpha \sum_{\alpha'} \mathbf{x}^{(\alpha)} \mathbf{x}^{(\alpha')} \Big(\sum_b w_b \mathbf{x}_b^{(\alpha)} \mathbf{x}_b^{(\alpha')} \Big) \\ &= \sum_\alpha \sum_{\alpha'} w_{\alpha\alpha'} \mathbf{x}^{(\alpha)} \mathbf{x}^{(\alpha')} \\ &= \mathbf{x}^T \mathbf{W} \mathbf{x} \end{aligned} \tag{2.87}$$

The obtained expression is equivalent to the second term in eq. 2.83. The advantage of it is that it no longer requires summation over basis vectors.

In practice, all available basis vectors can be used during the training phase and the eq. (2.87) in the prediction phase (during MD run) to cut computations. Note that obtained coefficient matrix $\mathbf{W}$ is symmetric which can be further exploited during the prediction stage.

2.5 Machine Learning Interatomic Potentials

2.5.1 Overview

The motivation behind *machine learning interatomic potentials* (MLIP) is to fill a gap between very expensive but highly accurate quantum modelling and cheap but often lacking in accuracy *classical interatomic potentials* (CIP as discussed in section 2.2).

Briefly, the purpose of MLIP is to represent the potential energy surface (PES) of a collection of atoms. The PES is a function of one or more coordinates and is differentiable with respect to to the atom coordinates. The latter property allows for computation of forces and consequently MD simulations.

The mathematical formulation of MLIP is in general more flexible as compared with CIP allowing to represent more complex PES. However, this comes at the cost as it does not take advantage of some physical insight which might by

obtained from the analysis of the system. Furthermore, MLIP usually do not perform well when extrapolating beyond the fitting region. The development of generalised MLIP is a very time consuming process and usually requires vasts amounts of training data. One might argue that instead of using physical insight to formulate functional form of the potential, this insight is spent during the development of a training data set suitable for the intended application.

MLIP are subject to the same constraints as their classical counterparts (see 2.2). One might try to directly use atomic coordinates as a feature vector and map it directly to the potential energy. However this does break some of the required invariances 2.2.0.4. Moreover such potential would be restricted to systems which are only identical to the training data sets. Any of those reasons alone is sufficient to rule out this naive approach.

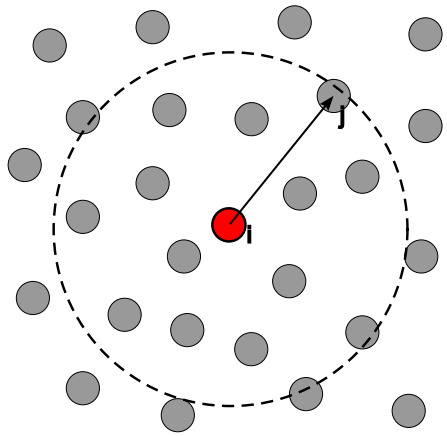


Figure 2.3: Local atomic environment of atom i.

An alternative approach is to first preprocess atomic coordinates into form which resolves aforementioned problems. Such preprocessed feature vectors are called *descriptors* or *fingerprints* and are discussed in detail in section 2.5.2. One popular choice is to construct descriptor vector for every atom in the system by considering its local atomic environment as shown in fig. 2.3.

Once descriptors are calculated the machine learning regression is employed to map them to some target values. As the goal of an MLIP is to represent the PES, the natural choice is to use potential energy as the target variable. The machine learning regression is discussed in section 2.5.3.

Once trained the machine learning interatomic potential is capable of predicting potential energy given some atomic configuration. MLIPs are often deployed to calculate energies and forces in the molecular dynamic simulation. This step usually requires coding a custom interface which links energy model to the MD package.

The mathematical formulation for MLIP follows ideas developed in section

2.2 where the total energy of the system of atoms is decomposed into local atomic contributions U_i. The machine learning approach follows the similar suit. Specifically, each local atomic environment within distance r_{cut} from the central atom i is represented by the local descriptor vector $\mathbf{d}_i$. The local energy is then obtained by feeding this descriptor into a trained MLIP model $\mathcal{E}$

$$U_{tot} = \sum_{i}^{N} U_i = \sum_{i}^{N} \mathcal{E}(\mathbf{d}_i) \tag{2.88}$$

where summations are over all atoms in the simulation.

The MLIPs have a defined functional form therefore interatomic forces can be obtained by applying the chain rule for differentiation

$$\mathbf{F}_j = -\frac{\partial U_{tot}}{\partial \mathbf{r}_j} = -\sum_{i}^{N} \frac{\partial \mathcal{E}(\mathbf{d}_i)}{\partial \mathbf{d}_i} \frac{\partial \mathbf{d}_i}{\partial \mathbf{r}_j} \tag{2.89}$$

here $\mathbf{F}_j$ is the force acting on atom j with coordinate $\mathbf{r}_j$.

Lastly, the virial stress tensor is obtained from [190]

$$\mathbf{S} = \sum_{j}^{N} \mathbf{r}_j \otimes \mathbf{F}_j \tag{2.90}$$

2.5.2 Descriptors

A descriptor is a specialised type of feature vector. The purpose of it is to represent either the system of atoms or the local environment of a particular atom within some cutoff distance. The latter approach is considered advantageous as it allows for the partitioning of the potential energy functional. The descriptor vector must satisfy a number of physical constraints to be valid such as

- Invariance with respect to permutation of atoms of the same species
- Invariance with respect to inversion, translation and rotation of the system

Those requirements have been discussed previously in section 2.2.0.4 and here are just restated for convenience. Moreover the descriptor function must be smooth and differentiable to allow calculation of forces without any unphysical

discontinuities. Ideally, the mapping between the atomic environment and descriptor is one-to-one (bijection)). The over-completeness of the set of descriptors may affect performance of the model. For example, in models using kernel ridge regression (2.4.4.3) this will increase the number of basis vectors required to make accurate prediction. Most importantly though both complete and overcomplete sets of descriptors provide unique description of the local atomic environment. The opposite case, where the same descriptor represents two different atomic environments, makes it impossible to fit the two configurations independently, which limits the capability of the model to describe reality. There are also other practical properties of the descriptor which should be considered.

- economy - how computationally expensive is it to evaluate the descriptor for a given atomic environment
- complexity - complex descriptors might require more training data
- scaling - how does descriptor (and model in general) perform when trained on incomplete training data which is the case in general
- physically grounded - in principle this should allow for better generality and scaling of the model when deployed in the region of the configurational space which is far away from where it was trained [114].

While it is possible to work with descriptors which represent the entire system of atoms (e.g. MD simulation box) they introduce number of unwanted restrictions. The transferability of such models between systems of atoms with varying number of elements is nonexistent. In principle such models work only for fixed number of atoms which is equal to the number of atoms used during training. Moreover, the development of training data base is restricted to small number of atoms in a box due to the high computational cost associated with high quality quantum mechanical computations such as DFT. Those two reasons alone are enough to consider such an approach suitable only for toy models.

The alternative and certainly more powerful approach is to leverage ideas used in the development of CIPs, namely the partitioning of the total energy function into local atomic contributions. Here, each atom, or more specifically its local environment, is represented by a descriptor. Therefore for each atom i there is a corresponding descriptor $\mathbf{d}_i$ representing its local atomic environment within a cutoff distance r_{cut} of this central atom. The cutoff distance is optimised

during the fitting process to balance model accuracy and speed. The premise here is that the forces are relatively short ranged and negligible beyond the r_{cut}. Obviously this methodology fails for systems where electrostatic long range interactions are important. However the ML model can be augmented with the separate evaluation of the long range interactions in the reciprocal space (Ewald summation) while the short range interactions are still computed with descriptors in the real space.

Since the inception of the field of MLIPs many descriptors have been proposed. The earliest descriptors can be dated back to late 1990s when simple models where developed for some low dimensional systems [18, 27, 68, 116]. The early ML models suffered from the lack of required physical invariances, or were fixed to the particular problem at hand, therefore being short of required generality. Those initial failings were, however, crucial to help recognise the challenge and pave the way for further developments in the field.

It is beyond the scope of this work to cover all developments in the field. It is worth noticing however that, similar to classical interatomic potentials, the descriptors can be classified as two-body, three-body and many body. The many body descriptors are further categorised based on their ability to capture angular dependence of their local atomic environment. Noteworthy is the fact that the ability (or lack thereof) of the descriptor to capture particular type of interaction within a system of atoms does not necessary limits its applicability. One must remember that by using nonlinear regression models the many-body nature of the resulting potential is recovered.

So far the discussion in this section was limited to monatomic systems. The extension for multi-species is possibly and number of schemes has been proposed. One of the first proposed solutions was to calculate separate descriptors for each pair of species [16]. So for binary system of species A and B three descriptors would be calculated, namely A-A, A-B and B-B. The final descriptor for the central atom i would be composed by concatenating those three preliminary vectors. While this solution is certainly valid and provides excellent ability to distinguish different atomic environments its main limitation is computational cost and scaling with the number of species in the system. Even for the system of three different atom types the computational cost is often excessive when paired with more sophisticated descriptors as the cost increases quadratically with the number of chemical species.

An alternative approach is to represent complex multi-component environments by the union of two descriptors [7] with constant complexity regardless of the number of atom types. The first descriptor represents the structure of the local atomic environment and is obtained in the same fashion as in the monatomic case by treating all atoms as being of the same chemical type. In other words the first descriptor captures the structure of the environment. The purpose of the second descriptor is to encode environment chemical composition. It is computed using the same functions as for structural descriptor but this time weighted by species dependent coefficients.

Yet, another approach is to compute weighted descriptors directly in the similar fashion to how this problem is usually solved with CIP [69]. Instead of computing a compositional descriptor as mentioned above, the method introduces implicit element-dependent weighting factors. Specifically, each interaction is weighted and the chemical environment information is now directly incorporated into the descriptor.

In the remainder of this section a number of descriptors are reviewed based on the aforementioned categorisation and, arguably, their popularity within the community.

2.5.2.1 Atom Centred Symmetry Functions

Pioneering work by Behler and Parrinello [19] introduced now widely popular atom centred descriptors known as *atom centred symmetry functions* (ACSF). ACSF can be categorised into two groups: radial or angular.

Radial descriptors are build out of Gaussians with two hyper parameters η and R_s which control width and position of the Gaussian respectively. Every descriptor is composed from a number of Gaussians of different width and positions, optimised for the given crystal structure(s).

Every component of the radial ACSF consists of a Gaussian and the cutoff function f_c. The purpose of the cutoff function is to ensure that a symmetry function and its derivative smoothly tend to zero at the cutoff distance R_c. The radial ACFS is defined as

$$G_i^2 = \sum_j e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{2.91}$$

where R_{ij} is the distance between atoms i and j and the cutoff function is

$$f_c(R_{ij}) = \begin{cases} 0.5\Big[\cos\left(\frac{\pi R_{ij}}{R_c} + 1\right)\Big] & \text{if } R_{ij} \leq R_c \\ 0 & \text{otherwise} \end{cases} \tag{2.92}$$

When radial descriptors are fitted to energies using ordinary linear regression using identity basis functions the effective potential is two-body in nature. The more sophisticated non-linear methods will generate many-body potentials with similar limitations encountered by the pair functionals (see s2.2.4.1) such as Finnis-Sinclair potentials. Namely, the assumption of spherically distributed charge density.

The angular descriptors account for the angular distribution of atoms and are better suited for systems where strong directional bonding is present. The angular descriptors are built by summing over triplets of atoms. Specifically, the sum is over all cosine values of the angle θ_{ijk} centred on atom i.

$$G_i^3(R_{ij}) = 2^{(1-\zeta)} \sum_{j,k\neq i} (1 + \lambda \cos\theta_{ijk})^{\zeta} e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{2.93}$$

where ζ, η and λ are adjustable hyper parameters. The computational cost of angular descriptors is high as it requires iteration over combinations of i-, j- and k-atoms.

It was considered for a while that ACSF are likely to form an overcomplete set in case the infinite series of the basis set expansion is used [10]. However, it has been shown that even models that utilises four-body correlations will incorrectly give identical results for different configurations [141].

In practice, the series is truncated to balance computational efficiency with required accuracy. It was demonstrated that it fails to uniquely distinguish configurations and the lower expansion limit [10]. Even though the representation improves when higher angular resolutions are included, the highly oscillating basis functions in this case will necessitate extensive training data sets.

2.5.2.2 SO(3) Power Spectrum - Smooth Overlap of Atomic Positions

In the seminal work on *Gaussian approximation potentials* (GAP) the descriptor describing the local atomic environment is build out of spherical harmonics [11].

To obtain such an expansion the local atomic density is first computed using delta functions. However it was recognised early that such an expansion would lead to numerical instabilities, as even a small change in atomic positions might result in large change in potential energy. Instead of delta functions an expansion using Gaussians was used resulting in smooth overlap of atomic positions descriptor (SOAP) [10]

$$\rho(\mathbf{r}) = \sum_i e^{(-\alpha|\mathbf{r}-\mathbf{r}_i|^2)} = \sum_i \sum_{nlm} c_{nlm} g_n(r) Y_{lm}(\hat{\mathbf{r}}) \tag{2.94}$$

where Y_{lm} denotes the Laplace's spherical harmonics, g_n is a radial basis function and coefficients c_{nml} are given by

$$c_{nml} = \langle Y_{lm} g_n(r) | \rho \rangle \tag{2.95}$$

where the integral above requires use of the modified spherical Bessel functions of the first kind. When an appropriate orthogonal radial basis function is chosen, the overlap between an atomic environment and its rotated counterpart can be obtained from

$$S(\rho, \hat{\mathrm{R}}\rho') = \int \rho(\mathbf{r}) \rho'(\hat{\mathrm{R}}\mathbf{r}) \mathrm{d}\mathbf{r} = \sum_{lmm'} J^l_{mm'} D^l_{mm'}(\hat{\mathrm{R}}) \tag{2.96}$$

where $D^l_{mm'}$ is a Wigner matrix and coefficients $J^l_{mm'}$ are given by

$$J_{nn'l} = \sum_m c_{nlm} (c_{n'lm})^* \tag{2.97}$$

and are equivalent to the power spectrum $p_{nn'l}$ as shown in [10].

Even though SOAP is not injective [141] it has been shown that it scales favourably in comparison with other popular descriptors [10]. In other words even at a relatively low level of expansion the descriptor is capable to uniquely distinguish training set configurations. It is also possible that inclusion of higher order correlations (four-body and above) may overcome this lack of injectivity [54].

2.5.2.3 Moment Tensor Descriptor - Moment Tensor Potentials

The basis functions used in the implementation of the moment tensor descriptor (MTD) [76, 78, 162] are polynomial in nature and similar to atomic cluster expansion [53] and related to the permutation invariant polynomial basis descriptors [196].

The MTD of the atomic environment of the ith atom, $\mathbf{n}_i$, is defined as

$$M_{\mu,\nu}(\mathbf{n}_i) = \sum_j f_\mu(r_{ij}) \underbrace{\mathbf{r}_{ij} \otimes \cdots \otimes \mathbf{r}_{ij}}_{\nu \text{ times}} \tag{2.98}$$

where f_μ is a radial part and $\underbrace{\mathbf{r}_{ij} \otimes \cdots \otimes \mathbf{r}_{ij}}_{\nu \text{ times}}$ contains angular information. The angular part is a tensor of rank ν as the outer product $\otimes$ is performed ν times. So for $\nu = 0$ is just a scalar, $\nu = 1$ gives a vector between atoms i and j and $\nu = 2$ is a symmetric matrix with its diagonal elements being simply squares of $\mathbf{r}_{ij}$ components and the remaining three off-diagonal terms are xy, xz and yz.

The radial part is given by

$$f_\mu(r_{ij}) = \sum_{B=1}^{N_Q} c_\mu^{(B)} Q^{(B)}(r_{ij}) \tag{2.99}$$

where $c_\mu^{(B)}$ is an expansion coefficient and the radial basis functions are composed of polynomial functions such as Chebyshev polynomials smoothed with a cutoff function. Such smoothing is common among different types of descriptors and ensures appropriate behaviour when ith atom neighbours move in and out of its local atomic environment.

To obtain a final descriptor vector which would be suitable for ML regression the MTDs are first contracted into scalars. The contraction is the process in which moments are collapsed into corresponding scalars using relevant inner product operation. Thanks to this process the obtained basis functions are invariant to atomic permutation, reflection and rotation.

To balance the accuracy of the angular expansion and also the computational efficiency the level of moment is first defined as

$$\text{lev} M_{\mu,\nu} = 2 + 4\mu + \nu \tag{2.100}$$

where the coefficients in the expansion above are empirically obtained from thorough testing performed in [77]. Second, the level of multiplication is calculated by summing over levels of the corresponding tensors. For example $\langle M_{1,2}, M_{0,2} \rangle$ is a Frobenius inner product (contracted scalar) with the corresponding level equal to 12.

The size of the descriptor vector obtained using MTD is therefore dependent on the number of radial basis functions and the number of level of moments used. By setting the maximum level restricts what moments are being used for the descriptor vector [134]. The moments as defined in eq. 2.98 have the following mechanical interpretation [162]. $M_{0,0}$ gives the number of atoms within the cutoff distance, $M_{0,1}$ is the their centre of mass and $M_{0,2}$ is the second moment of inertia. The third moment can loosely be interpreted as a measure of asymmetry in the distribution of atoms.

2.5.3 Regression

Regression is a process of finding the relationship between the independent variables (atomic coordinates) and the scalar response such as potential energy. This relation is certainly nontrivial and additionally complicated by the physically imposed requirements 2.2.0.4. During the fitting procedure, this complex relationship is being captured partially by the choice of the descriptor. The main idea is that appropriate selection of the model will complement the descriptor, allowing sufficient representation of the system. In other words, the nonlinear physics of the problem may be mostly encapsulated by the choice of descriptor, or by the choice of regression model.

In this section I will briefly review popular regression methods suitable for the purpose of the development of the interatomic potentials. All regression models presented below belong to the class of supervised machine learning methods.

The *linear regression* discussed in section 2.4.4.1 provides the simplest and fastest fitting procedure. Note it is still able to represent nonlinear nature of the potential energy function if this nonlinearity can be incorporated into the descriptor of choice. Another advantage is that if a mathematical form of the descriptor is physically informed, then the linear model might generalise much better as compared with more complex models.

Feed forward neural networks (NN) [131] provides functional flexibility to represent complex functions. NN are built of layers each containing a number of nodes. Each node is interconnected with all other nodes in the following layer. The last layer acts as a lens which converges its input onto the final result - the operation equivalent to the linear regression. The strength of connections between nodes is modulated by weight parameters which are fitted during the training process. When a descriptor is presented to the NN, its components propagate through the network in a nonlinear fashion until merged by the last layer.

A *Gaussian process regression* model (GPR) is a conditional probability distribution over all possible functions that fit a set of training data points [148]. This is in contrast to the usual approach where only one nonlinear function is expected to fit the data set. GPR allow for a possibility that there might be more than one function that fits the training data equally well. In consequence, the conditional posterior distribution provides measure of uncertainty. The prior knowledge about the modelled function, which in our case is PES, is incorporated in terms of kernels also known as a covariance function. The kernel provides the similarity measure between two different descriptors. The prior distribution of infinite functions in GPR is a multivariate normal distribution. The prior provides the expected output for the function before observing any data points. Once model is presented with training data, every covariance function will have a value associated with it. A prediction at an unknown point can be obtained by combining a prior with a likelihood function for the observed values. One can think of it in a sequential manner: after observing the new data point the prior becomes the posterior. In the next iteration the current posterior is used as a prior and so on. In principle, every new data entry not only improves model predicting power but also refines our understanding of its performance.

2.6 Molecular Crystal Phases

The ability of solid to exist in more than one crystal structure is often referred as polymorphism or allotropy for pure chemical elements. In general, crystals can be classified as atomic, molecular or a mixture of both. The arrangement of atoms in crystalline solids can be described using space groups also known as crystallographic groups.

A space group is the symmetry group which involves operations such as rotations, reflections, translation screw dislocations and glide reflections. Therefore a space group associated with a given polymorph provides a list of symmetry operations which preserves crystal invariance. In practice it allows to uniquely describe the positions of the atoms in the system.

The molecular crystal phases differ from crystal formed from pure elements as molecules can have additional degrees of freedom. For example the molecule might freely rotate around the fixed point in the crystal.

For a given P,T condition only one phase can be considered as stable at equilibrium - the one with the lowest free energy. For some systems, in addition to a lowest energy phase a number of competing phases can be observed. Those are referred as metastable phases. It is often the case that their energies differ a little from the lowest energy structure but the transition pathway involves high energy barrier. The metastability phenomena manifest itself in both experiments and simulations.

Depending on P-T path taken from the initial structure different phases can be observed at the final P, T conditions. In the experimental settings it is usually impossible to determine in such a case which phase is the most stable. To mitigate some of these challenges experiments often involve slow changes in P, T and long equilibration times.

In atomistic simulations determination of ground state structure is fairly straightforward as absolute enthalpies are readily available. However, at elevated temperatures the situation is more complex. The short timescales might not allow system to jump over high energy barriers. Moreover the shape and size of the simulation box might hinder the phase transitions. In principle computation of Gibbs free energy is possible with thermodynamic integration but in practice is very laborious and computationally expensive.

Chapter 3

Machine learning library for interatomic potentials

The main motivation behind the development of the new, community driven, software and library (the code) is the ability to design novel machine learning interatomic potentials and also to allow rapid deployment of those in the large-scale atomistic simulations environment.

In addition, the code aims to minimise practitioner's effort during the potential development stage by introducing iterative two-stage fitting process inspired by the accomplishments in the field of classical interatomic potentials. The details of this optimisation procedure are provided in chapter 4 along with some examples.

At the moment, the code provides a number of popular atom-centred descriptors and two regression methods: Bayesian Linear Regression (BLR) and Kernel Ridge Regression (KRR) Most importantly, its object-oriented design allows for a quick development of new ideas such as adding new descriptors or regression models. The code is fully interfaced with the LAMMPS package [189] by native C++ plugin such that new code developments are immediately available for the deployment in the simulation.

The code provides an easy to use command line interface (CLI) along the more advanced option to be utilised as an C++ library. The code is open-sourced and available for download from

`https://git.ecdf.ed.ac.uk/s1351949/ta-dah`.

The documentation along with the user guide and some usage examples are available at

`https://ta-dah.readthedocs.io/en/latest/`.

While a relatively large number of codes have been developed to support the invention of MLIPs they are mostly limited either by providing single type of descriptor and regression method or can be considered as toy models because of their inefficient implementation or lack of interface to MD software. Besides, a disproportionately large proportion of packages supports neural networks in comparison to linear models such as KRR and BLR.

The code is coined "*Ta-dah!* " to emphasise fun, amazement, joy and perhaps a little bit of witchcraft which is associated with the development of interatomic potentials. Building an interatomic potential is like pulling a rabbit out from the hat. It is magic but it is not.

3.1 Software Technical Overview

The guiding principles during the design stage of the code were flexibility of usage, computational efficiency and extensibility of the code base. In this section a brief overview on how aforementioned principles where achieved from a technical perspective.

The code is written in modern C++ language in accordance with the standard ISO/IEC 14882:2011[1]. Standardisation of C++ under ISO guarantees its longevity as future compilers are expected to support old code. As a compiled language it is computationally efficient and can be optimised for usage across many platforms. As it stands the code has been tested on various Linux distributions including HPE Cray Linux Environment used by Archer2 supercomputing system.

The software is build on a number of excellent open-sourced libraries. Notably, linear algebra operations are performed by Eigen [79]. Hyper parameter optimiser is possible thanks to MaxLIPO+TR algorithm from Dlib [103] and requires LAMMPS to be available (compiled) as a shared library. Apart from LAMMPS,

[1]Informally known as C++11

the code does not require installation of those components per se. Building and compilation of the code is greatly simplified as is governed by CMake. The code comes with a limited set of unit tests which is progressively being expanded.

The object oriented paradigm is used to organise the code base and allow for new components to be added with the minimal effort. For example, adding a new descriptor to the software would require implementation of a single class according to the specification. Once implemented a new descriptor not only seamlessly works with the rest of the code, such as regression classes, but is also immediately available within LAMMPS for simulations.

To allow flexible usage of the code as a stand-alone library elements of the generic programming are employed such as class templates. This paradigm allows to write general algorithms which can work with various data structures without compromising efficiency. Specifically, it provides compile-time polymorphism required for the library and also improves code re-usability by generalising software components. The flexibility provided by templates is of particular importance in a fast-changing field such as development of methods for MLIPs.

In contrast to the library mode, the command line interface and LAMMPS plugin require run-time polymorphism - namely, selection of model components based on a configuration file. Writing the same code twice is obviously a poor solution. Instead a factory method is employed which permits coexistence of run-time along the generic programming compile-time polymorphism. The factory method is an object oriented design pattern which allows for construction of product classes without defining concrete ones [66].

The code provides a generic plugin interface for LAMMPS package. Since both pieces of software are written in C++ there is no additional overhead associated with translation of data structures between two different programming languages. The interface separates the data structures where appropriate and the computation of descriptors and prediction of energies and forces from the simulation software. For example, LAMMPS computes nearest-neighbour lists and requests ML code to return the force for a given type of interaction. This object-oriented solution improves on extensibility of the code without sacrificing its efficiency.

Computationally demanding tasks are parallelised with hybrid Open Multi-Processing (OpenMP) and Message Passing Interface (MPI) when appropriate, such as calculation of descriptors during the training stage or the optimisation of

hyper parameters. Besides, the code still works seamlessly with LAMMPS which supports native (MPI). The current implementation does not parallelise routines responsible for solving linear systems of equations due to limitations of the Eigen library [79]. We plan to replace those with ScalaPACK implementation in the near future.

3.2 Capabilities

The code is intended to maximise the flexibility of a user during the development stage while at the same time ensuring high computational efficiency. The logic behind this philosophy is quite simple. The development of machine learning models happens in an iterative manner: build a model, test it, repeat. The less time that is spent waiting for a computation to finish, the more time is available for more meaningful tasks. Moreover, the efficient training process allows for development of a hyper parameters optimisation method described in chapter 4

In a nutshell, the code provides capabilities to train models using various descriptors, cutoffs and regression models. Once trained the model can be deployed to perform MD simulations with LAMMPS or to predict energy, virial stresses and forces for a given configuration(s) of atoms. The model is capable to provide uncertainties on its predictions, such as energy and forces, as well as uncertainty on learned weight coefficients. The development of models is streamlined by introducing hyper parameter optimiser. The code comes with a simple analytical module which provides basic statistics during model fitting, such as RMSE on predicted forces.

In the remaining part of this section training, prediction and simulation with LAMMPS are discussed in more detail along with some main features of the code. The hyper parameter optimiser is then introduced with some basic examples to illustrate its potential.

3.2.1 Training procedure

The training procedure, in principle, should begin by specifying requirements on the model. In particular, questions such as what is the intended application for the potential and in consequence the number of atoms the model is suppose to

handle? What region of the configurational phase space is relevant? Is potential being developed for narrow specialised application or perhaps it is designed as a general purpose which is conceivably a much more challenging task?

The next step is to obtain high quality data for fitting. The data should cover configurational space required by the intended application but should also be sparse enough so no unnecessary calculations are begin performed. Once QM data are available they should be converted to a specific format required by the code.

The model parameters are then specified in the configurational file. The more advanced option is to provide them directly in the C++ code in the case when library mode is used. However, the configurational file typically provides sufficient flexibility for both CLI and library mode. Apart from some minor parameters, which are covered in the documentation, the critical model decisions are: Is two-body descriptor and/or many-body descriptor are being used? If yes, specify their types and corresponding cutoff functions and cutoff distances. Most descriptors require a set of hyper parameters (HPs) to be provided. Next, choose model such as BLR or KRR and specify corresponding basis functions or kernels respectively. Is training being performed on energies only or perhaps forces and virial stress will be used as well. To control overfitting the regularisation parameter can be specified directly or the code can estimate it using evidence approximation algorithm. It is sometime advantageous to standardise descriptors feature wise by subtracting their mean and dividing by the standard deviation. Finally, general factors can be set to weight energies, forces, and stresses. It is also possible to set those weights directly in the data set file for every structure. If no weight is set in the data file it will default to unity.

3.2.1.1 Training Database

A training database consists of one or more training set files. Every set file must contain at least one structure. For our purpose, a structure contains matrix of lattice vectors, stress tensor, set of atomic configurations and potential energy. Every atom in a structure is identifiable by its chemical element name, position and force acting on it. The format of the set file is explained in the documentation. The parser for a set file allow varying number of structures each containing different number of atoms in a simulation box.

The training database is usually generated using higher order quantum mechanical theory such as density functional theory or post-Hartree–Fock ab initio quantum chemistry method such as coupled clusters. The code does not provide any tools for the generation of the database. It is usually a straightforward process to convert quantum mechanical computations into a suitable format.

While in principle the code is unit agnostic, it has only been tested with the following units: distance in Å, energy in eV, force in ev/Å, pressure in GPa. We note that the conversion between units is a rather elementary task.

The training database should cover configurational space which is relevant to the intended model application. ML models are well known for their poor predicting power outside their applicability domain.

3.2.1.2 Descriptors

Perhaps the most critical choice for the MLIP model is a selection of descriptors and corresponding cutoff functions. The code supports computation of two-body and many-body descriptors. The resulting final descriptor which will be passed to a regression model can consist of any combination of those, i.e., plain two-body, many-body or a concatenation of those two.

Every type of descriptor can be matched with a suitable cutoff function and a custom cutoff distance. The dummy cutoff function is also provided for cases where descriptor functional form smoothly goes to zero at the cutoff distance. The functional forms of descriptors, supported by the code, are given in 3.3.1.

Some of descriptors can be considered as *classical* such as simple Lennard-Jones or Mie type descriptors (see documentation for complete list of available descriptors). While those are clearly limited in their functional form and one might argue that it defeats the purpose of ML potentials where flexible mathematical function is postulated. Nevertheless they are made available as a bridge between CIP and MLIP and such they allow for a direct comparison between different type of models. Moreover they provide a simple baseline to which more sophisticated potentials can be compared with.

Three-body (and higher) descriptors are implemented as many-body. For example, in the embedded atom descriptor [183, 211] when charge density is being expanded in terms of angular resolved functions, this is the case.

3.2.1.3 Model selection and regression

Both implemented regression models, namely BLR and KRR, require selection of a corresponding basis function or kernel respectively. By selecting identity basis functions or kernel the resulting model is equivalent to regularised linear regression. Other choices will lead to non-linear models. The model should complement selected descriptor(s) such that the essential physics of the problem is captured, e.g. two-body descriptor and linear regression model result in a simple pairwise model such as Lennard-Jones. However the same descriptor with a non-linear regression model will result in a many-body potential, which could be equivalent to FS or EAM. The regularisation parameter in eq. 2.66 can be provided by hand or estimated with the evidence approximation algorithm [26].

The training process can be performed using combination of structure energies, atomic forces and virial stresses. The importance of those can by weighted by either global scale factors in the configuration file or individually by providing scale factors for selected configurations. The training process happens in a closed form providing optimal solution to a regression problem.

The final output of a training process is a **pot.tadah** potential file which can be either directly used for a prediction or LAMMPS simulation using provided interface. The code can also provide uncertainties on regression coefficients which can by used to optimise hyper parameters or identify relevant training database structure.

3.2.2 Prediction procedure

The prediction process can be achieved using CLI or a custom written C++ script. The format of the prediction data set is the same as used during the training process (3.2.1.1). The data set must contain at least lattice cell vectors and atomic species and positions. Energy, virial stresses and forces are optional and only needed when statistics on prediction are required. To satisfy the data set parser, unavailable quantities should be set to zero.

The code is capable of predicting energy per atom, atomic forces and virial stresses given a potential file. The corresponding output is written to **energy.pred**, **forces.pred**, **stress.pred** files respectively. The format of those files is explained in the code documentation. The model is capable of estimating the uncertainty

of its predictions without knowing target values.

3.2.3 LAMMPS interface

The code provides a custom LAMMPS interface which is distributed as an optional LAMMPS package. Thanks to an efficient implementation the code allows for a simulation of large-scale systems. In principle, simple linear models should allow for a molecular dynamics simulation of millions of atoms on modern high performance computing facilities such as Archer2. However, we note that more complex models can be orders of magnitude slower.

The interface is currently not distributed with LAMMPS and has to be copied from the code to the LAMMPS directory. See LAMMPS documentation on optional packages for more details. Once LAMMPS is compiled with the code interface the potential files can be used in the usual way, e.g.

```
pair_style      tadah/tadah
pair_coeff      * * pot.tadah Ta
```

3.3 Essential theory

This section covers theory upon which the code is based on and is required for the development of ML potentials. The generalised functional form of two- and many-body local atomic descriptors are presented. The fully localised *blip* functions are introduced which can be used in place of Gaussian functions utilised by various descriptors. The theory behind Bayesian Linear Regression and Kernel Ridge Regression is covered in sections 2.4.4.2 and 2.4.4.3 respectively with an emphasis on the former as it is often our method of choice.

3.3.1 Generalised Descriptors

The code provides two types of descriptors: two-body and many-body. Most descriptors are specified by a set of hyper parameters. In ML methodology a hyper parameter is a parameter which controls the learning process but it is not obtained during the training process. In case of descriptors, hyper parameters can, for example, control positions and widths of Gaussians or can be exponents

in the Mie type descriptor (the generalised case of the Lennard-Jones descriptor).

Henceforth, we define a set of hyper parameters which completely specify a descriptor vector as

$$\{\{\zeta_p^1, \dots\}_p\}_{p=1}^{N_p} = \{\{\zeta_1^1, \dots\}_1, \{\zeta_2^1, \dots\}_2, \dots, \{\zeta_{N_p}^1, \dots\}_{N_p}\} \tag{3.1}$$

such that $\{\zeta_p^1, \dots\}_p$ is a subset of hyper parameters for the p-th component of the descriptor vector which will be shorthanded as $\{\zeta\}_p$ from now on. A subset contains varying number of hyper parameters (indicated by $\dots$) which are relevant to a functional form of a descriptor.

3.3.1.1 Two-body descriptor

The functional form for the p-th component of of the two-body descriptor of the ith atom with hyper parameters $\{\zeta_p^1, \dots\}_p$, is

$$v_p^{(i)} = \sum_{j \neq i} \alpha_{ij} B^{\{\zeta\}_p}(r_{ij}) f_c^{\{\zeta\}_p}(r_{ij}) \tag{3.2}$$

where B is a function with parameters $\{\zeta\}_p$ such as Gaussian function in ACSF (eq. 2.91), f_c is a cutoff function and r_{ij} is a distance between atoms i and j. The summation is over all neighbouring atoms of central atom i which are within the cutoff distance r_c. Note that r_c is included in the $\{\zeta\}_p$ subset of hyper parameters as it is might by required by the B functions as well.

The hyper parameter α_{ij} modulates the strength of an interaction between two atoms of the same or different species. It is similar to the weighting introduced in *weighted atom-centred symmetry functions* [69] but our implementation is symmetric allowing exploitation of so-called *half neighbour lists*, i.e., just one computation of descriptor is required for every i-j interaction since pairwise forces are equal and opposite. Note that linear model used with wACSF results in $f_{ij} \neq f_{ji}$ for *half neighbour lists* which is unphysical, however we recognise that it is not the case when *full neighbour lists* are used. The code provides default value of $\alpha_{ij} = Z_i + Z_j$ where Z is an atomic number. In practice, it is usually beneficial to optimise this hyper parameter either by hand or using HPO (see 4).

3.3.1.2 Many-body descriptor

The many-body descriptors of the ith atom satisfy the following form

$$v_p^{(i)} = \mathcal{D}^{\{\zeta\}_p}\big(\boldsymbol{\rho}_i^{\{\zeta\}_p}\big) \tag{3.3}$$

where the vector of electron density $\boldsymbol{\rho}_i$ of the ith atom is build out by expanding the density using the basis set of choice.

$$\boldsymbol{\rho}_i^{\{\zeta\}_p} = \Big(\sum_j \alpha_j \psi_1^{\{\zeta\}_p}(\mathbf{r}_{ij}), \ldots, \sum_j \alpha_j \psi_{max}^{\{\zeta\}_p}(\mathbf{r}_{ij})\Big) \tag{3.4}$$

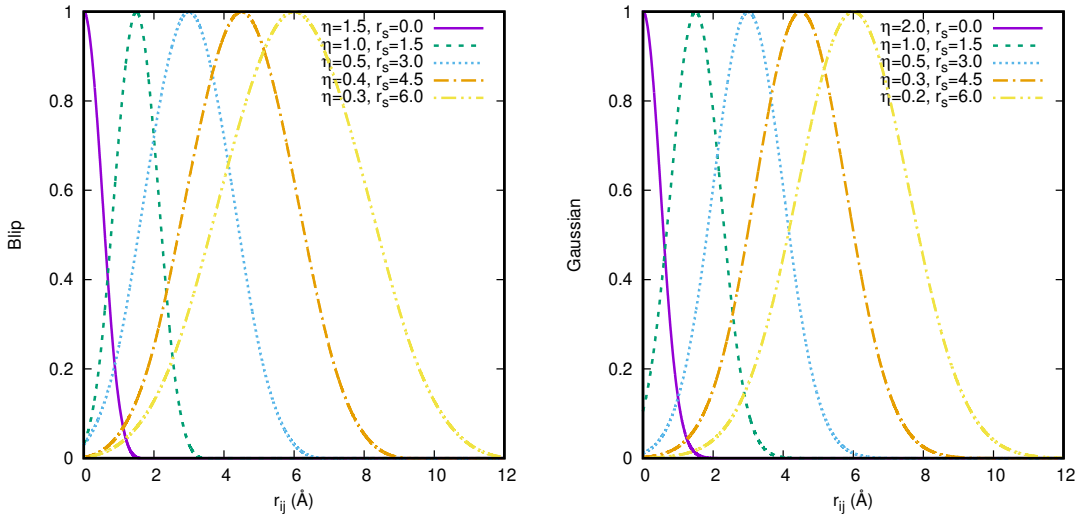
In case where only distance r_{ij} is used instead of the $\mathbf{r}_{ij}$ vector than the uniform density approximation is recovered and the linear model built out of this descriptor can be considered as generalisation of EAM- or FS potentials [47, 48, 64]. However when functions ψ are angular dependent such models generalise the idea behind modified EAM [12] models.

Note that the functional $\mathcal{D}$ must satisfy usual invariances (as discussed in 2.2.0.4) for linear models while in principle this invariance can be introduced by the non-linear model. In particular, the individual components of the density vector (eq. 3.4) can violate rotational invariance provided that they are then combined in such a way that the rotational invariance is preserved.

3.3.1.3 Blip functions

Many different types of descriptors in the literature make use of Gaussian functions. Historically, computation of Gaussian functions took hundreds of clock cycles on older processors, however this not the case any more as all modern CPUs have fast floating-point hardware. Perhaps a more relevant *unwanted feature* of a Gaussian, for the development of interatomic potentials, is their infinite span, i.e. the computation of a Gaussian is always required.

Here we introduce a simple substitution for Gaussians based on an earlier work on localised basis functions for the purpose of first-principle calculations [87]. The *blip functions* are centred on the points of a grid similarly to a Gaussian function. Their *shape* is similar to a Gaussian, but contrary to the former they completely vanish outside their limited domain. In other words they are fully



(a) Example grid made of blips.

(b) Example grid made of Gaussians.

Figure 3.1: An example grid of blip functions composed of B-splines along a similar grid build out of Gaussian functions. The two grids are not meant to be identical.

localised functions.

The *blip* function is composed piecewise out of B-spline polynomials in the four intervals [-2,-1], [-1,0], [0,1] and [1,2]. B-splines are localised basis functions used to represent functions in terms of cubic splines [156]. The *blip* function is defined for our purpose as

$$B(r) = \begin{cases} 1 - \frac{3}{2}r^2 + \frac{3}{4}|r|^3 & \text{if} \quad 0 < |r| < 1 \\ \frac{1}{4}(2 - |r|)^3 & \text{if} \quad 1 < |r| < 2 \\ 0 & \text{if} \quad |r| > 2 \end{cases} \tag{3.5}$$

where $r = \eta(r_{ij} - r_s)$ and r_s is a parameter which centres function on a grid position and η controls its shape such that $\eta/4$ is a span of a *blip*.

The *blip* function vanish smoothly to zero outside their specified interval. The function is continuous everywhere and so its first two derivatives. An example grid of *blip* functions along similar grid of Gaussians is illustrated in fig 3.1.

As mentioned previously a computation of Gaussian functions is nowadays implemented on the CPU hardware level resulting in high performance. However we find that the simple piecewise polynomial functions outperform those due to the fact that they are fully localised. The gain in performance does not come

from faster *blip* computations of B-spline functions but from the fact that less blips is usually calculated because of their localised span. This is particularly pronounced when the model grid contains many narrow functions instead of a few broader ones.

3.4 Example usage

The code below will train the linear model using two- and many- body descriptors.

```
// Read configuration file
Config config("config");
// Load training data
StructureDB stdb(config);
// Find nearest neighbours
NNFinder nnf(config);
nnf.calc(stdb);
// Define descriptors and cutoff functions
using D2=D2_BP;
using D3=D3_Dummy;
using DM=DM_EAD;
using C2 = Cut_Cos
using C3 = Cut_Dummy
using CM = Cut_Poly2
// Instantise calculcator for descriptors
DescriptorsCalc<D2,D3,DM,C2,C3,CM> dc(config);
// Instantise model and train
M_BLR<BF_Linear> model(config);
model.train(stdb,dc);
```

where the configuration file `config` contains information about cutoff distances and relevant hyper parameter settings.

Equivalently the same result can be obtained directly using the command line interface.

```
$ ta-dah train -c config
```

where the configurational file `config` specify the entire model completely and the result is a `pot.tadah` file.

The resulting potential file can be used for prediction on different data sets:

```
$ ta-dah predict -d dataset1 dataset2 -p pot.tadah -FSa
```

where `-FSa` flags indicate to compute forces, stresses and also provide basic analytics such as RMSE on former quantities.

3.5 Future development ideas

Ta-dah! has been developed with an intention not only to be easy to use but most importantly to allow extending the code base with new methods and features. The second requirement is particularly important is the fast changing field of MLIPs.

Below, is a list of features which are not included at the time of writing but are at least worth consideration.

- Python interface to C++ library
- Integration into Atomic Simulation Environment
- Regression with Neural Networks
- Active Learning
- On-the-fly MLIPs
- Support for automatic differentiation of models and descriptors
- Different algorithms for hyper parameter optimisation such as genetic algorithm (4.2) or particle swarm optimisation (4.2)
- Meta-programming approach to join descriptors of the same type, such as two different two-body descriptors into one.
- More descriptors...

Chapter 4

Two-Stage Fitting Procedure

In general, a ML model consists of two types of parameters: *learned parameters* (LP or simply *parameters*) which are obtained during the learning stage and *hyper parameters* (HP) which are preselected before the learning process commences. The choice of MLIPs architecture, such as ML algorithm and descriptors, represents our prior knowledge about the type of system that is being studied. The optimal model architecture is then achieved by tuning its HPs such as positions and widths of Gaussians in ACSF descriptor (eq. 2.91) or a topology of a NN. Even decisions about which regression algorithm to choose are considered as a HP selection problem. Typically, HPs cannot be established from the data alone during the LPs regression stage and must be set beforehand. Perhaps, the most common, complex and time consuming adjustment to the model performance is optimisation of model's HPs.

The development of ML models happens in an iterative manner and MLIPs are no exception. The common strategy employed when building ML models is to fit a model on the training data set and fine-tune HPs with the validation set [82]. The process is repeated until the objective function is either minimised in case of a loss function or maximised when a fitness function is under consideration. Testing on the validation set guides further development of the model and is used to estimate prediction error for model selection. The final model performance is then evaluated on the hold-out set.

For the development of MLIPs the above strategy is too simplistic and resource wasteful.

First, the final model performance cannot be evaluated on the test data alone - good match for energies, forces and stresses does not guarantee that the potential will perform well in MD simulations. The MLIP performance is related to the quality of the training data (configurational space coverage more generally) and to the ability of the model not only to interpolate between data points but also to extrapolate to unknown regions. The latter is strongly influenced by the model architecture.

Second, it is too expensive as one cannot afford to generate high quality large data sets using higher order quantum mechanical theory and then split it into usual 50% training, 25% validation and 25% test subsets [82]. The test set is usually drawn from the same distribution as the training data and given strong interpolating capabilities of ML models it is often the case that estimate of predictive accuracy of the model is overinflated.

Ideally, the fitting procedure would consists of evaluation step against the complex quantities such as a melting curve or various crystal phase stabilises at finite temperature. However, the accurate computation of the melt curve is a time consuming process while the latter cannot by obtained from the ensemble averages as Gibbs free energy is not a property of a microstate.

It is hopefully clear by now, that in the development of MLIPs, the model architecture as well as the training data should be considered as adjustable variables. Here we tackle optimisation of a model architecture while the optimisation of a training data set is beyond the scope of this chapter. The optimisation of the training data depends sensitively on the system and intended application.

The critical questions which guide our developments are: *What is the form of a global objective function which maximises model's performance? How to optimise this function?* and last but not least, *How to quantify MLIP performance itself?*

Here by the *the global objective function* we mean what is commonly known in the ML literature as *an evaluation function*. An evaluation function is essentially a set of metrics which are used to judge model's performance. The very basic test is an evaluation on the hold out set which is usually unavailable when developing an interatomic potential. Perhaps better metrics are model's ability to reproduce experimental pressure-volume curve or vacancy formation energy. *The global objective function* term is used specifically to indicate that it is, in fact, an optimisation problem.

The structure of this chapter is as follows. First, a brief overview of HPs selection process is given followed by a review of some common optimisation techniques. The new method is then presented for the optimisation of HPs during the development of MLIPs. The aim of the new method is to reduce the time required to construct a new potential, minimise overfitting of HPs while maximising model's accuracy as compared with QM calculations. Moreover, the method is intended to support the development of potentials which are suitable for the general purpose even in presence of suboptimal training data sets. To validate this new approach, it is deployed to a number of problems with increased complexity. The method is fully implemented into *Ta-dah!* package (chapter 3).

The presented approach is inspired by methods developed in [86]. Therein, a ML algorithm based on symbolic regression is used to construct MLIP using QM training data. A hypothesis space is then generated for physically meaningful expressions, such as LJ (2.2.1.1) and BOP (2.2.4.2). This search space is than explored by the genetic algorithm (4.2) to model natural selection process. The generated new functional forms for a potentials are then fitted to training data set energies in the usual way.

Herein, the approach is different. Instead of constraining the functional form from the limited hypothesis space we allow our model to explore mathematically more flexible ML descriptors. In principle the aim is the same, that is to obtain a functional form for the potential which is physically informed and can generalise well beyond the training data set. The approach is probabilistic in nature: Rather then choosing the best fitting function to a set of training data, we consider a number of functions which are close to the best fitting model. To do that we enforce additional constraints on a set of hypothetically valid functions, that is, those that reproduce training energies well. The physical constraint are represented by the global objective function which is than optimised. The details of the method are presented in section 4.3.

4.1 Background

The aim of HP optimisation is to minimise the loss function $\mathcal{L}(\mathcal{M}, \mathcal{T})$, where $\mathcal{M}$ is a model and $\mathcal{T}$ is a training data set

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \mathcal{L}(\mathcal{M}, \mathcal{T}) \tag{4.1}$$

where the model is parameterised by a set of HPs λ. The goal is to find a set of λ^* within a search space Λ such that desired model is obtained [117]

The optimisation of HPs differs from other optimisation problems [118]. Theoretically it should be possible to obtain the gradient of the objective function with respect to the model HPs. In practice this is rarely the case. One reason is because the search surface usually contains discontinuities [173] or is simply not differentiable in case of categorical HPs. The domain of HPs might either be continuous (e.g. regularisation parameter), discrete (e.g. number of Gaussians in the descriptor), binary (e.g whether to normalise descriptors) or categorical (e.g. kernel type) [206]. The domains of continuous and discrete HPs are usually bounded for practical purposes [24].

As with any ML optimisation technique one must control the risk of overfitting HPs to a given data set. In general, different data sets will have different optimal HPs values. This is a typical problem when HPs are optimised by hand without test set but can also happen when automated methods are employed. For example, naive optimisation of the degree of polynomial of BLR basis functions will usually lead to highly overfitted function. This problem is discussed further in 4.3 in relation to the optimisation of HPs for MLIP.

The process of optimising HPs consists of the following intertwined components: the ML regression algorithm and the corresponding loss function. The optimisation algorithm which will search the configurational space of HPs and use the evaluation function to measure models performance given different sets of HPs.

It has been shown previously that a genetic algorithm (4.2) (GA) can effectively search HPs configurational space for both classical and machine learning interatomic potentials [38, 39, 106, 113, 136, 161, 216]. Apart from GA ((4.2)), trial and error, also known as manual, fine-tuning of HPs is very competitive, as compared with automated methods, albeit extremely time consuming (see discussion in 4.2).

The ability of MLIP to accurately predict validation (or even better a hold out) set energies, forces and stresses is a necessary but not sufficient condition. The further testing is required to establish its performance in true case scenarios such as MD simulations. Testing of the MLIP candidates can be very laborious and usually a number of simulations have to be performed to establish its true performance. Moreover, the transferability of MLIPs is a well known limitation and is strongly correlated with the quality of the training data set. In general,

MLIPs are capable of high accuracy when constructing PES for local atomic configurations which are similar to those in the training data set. This is the consequence of using generic functions with large number of free parameters [72].

4.2 Hyper Parameter Optimisation Techniques

In this section a number of popular methods for hyper parameter optimisation is briefly examined and in section 4.3 the new procedure is developed for the specific needs of MLIPs. For a more comprehensive review the reader is referred to [118, 206].

Manual tuning of HPs is often a method of choice among ML practitioners. It is a simple *trial and error* algorithm[1] which is terminated when satisfactory results are obtained or one simply runs out of time. The success of this method relies on good understanding about a problem at hand and detailed comprehension of algorithms involved. Every iteration in the training process brings new knowledge which can be used to further fine-tune HPs.

The trial and error method is time consuming, in particular when model retraining takes significant time, but the improvements are often surprisingly large. However, for problems where the model contains large number of HPs (such as in MLIP) or the interaction between them is nonlinear further complicates the challenge. Another disadvantage of the manual method is that it makes research often non-reproducible and comparison between different models is more problematic - as it is often unclear whether the gain in performance is due to the better method or simply more diligent optimisation of HPs. The automated method for HPs optimisation attempts to alleviate at least some of those issues.

Grid search (GS) is a common brute force method employed in many optimisation tasks [93]. It is well suited for problems where configurational space is discrete and of low dimensionality. The GS method attempts every possible combination of HPs from the search space. The method can be easily parallelised however it suffers from *the curse of dimensionality* - computational cost increases exponentially with the number of dimensions. The common strategy when employing GS is to perform an initial search over coarse configurational space followed by the finer one for the most promising regions. The procedure can be

[1] Also known as Grad Steepest Descent - the optimal model parameters are obtained when the deadline is reached.

repeated number of times until convergence criteria is met or one simply run out of allocated time.

Random Search (RS) is another popular choice for the selection of HPs. Given enough time RS is capable of discovering global minimum, however, this is unlikely for high dimensional configurational spaces. RS samples combinations of HPs from predefined bounded distribution. RS can explore larger HP space as compared with GS [25]. Similarly to GS, the RS method is easy to implement and parallelise well as every evaluation is independent of the past.

The major limitation of both GS and RS is the fact that they do not use the information from the previous step to guide their next move. Therefore they often spend large amounts of time exploring useless areas of the search space. The simple manual search takes this into consideration when applied thoughtfully. The following automated algorithms attempts to amend this shortcoming.

Gradient descent (GD) algorithms work by numerically calculating the derivatives of the HPs search space [23]. The direction of the next move is based on the steepest gradient found. The point of calculation is selected at random and the computation might be repeated on a set of points. For high-dimensional optimisation problems a stochastic version is often employed where the actual computation of the gradient is replaced by its approximation. The GD works only on differentiable functions (or at least differentiable subset) which is often not the case for HP spaces. Moreover, the global minima is only reached for a convex functions while their efficiency drops significantly for non-convex ones.

Bayesian Optimisation (BO) [170] attempts to obtain a minimum in a high-dimensional HPs space by constructing a probabilistic model based on observed evidence. Contrary to GD, BO uses all previously acquired information instead of just a local gradient to decide its next move. The BO procedure usually involves building a surrogate model on currently available data, and then use it to integrate out uncertainty and build new predictive distribution. BO algorithm is computationally more expensive than GD, however, it should be able to reach minimum of complex functions in relatively fewer steps. The BO model is composed of two parts. First, the prior over functions to be optimised, such as Gaussian process prior [148, 159]. The prior represents assumptions about the objective function. Other popular surrogate models include random forest [92] and three Parzen estimator [24]. Second, the acquisition function which is used to decide the next move. The utility of the acquisition function is to balance

exploration of the unknown HP space against the exploitation of the promising regions where the minimum is most likely to occur. Ironically, BO models contain their own HPs, which strongly influence their effectiveness and can be difficult to establish. In some case BO model can have more HPs than the ML model which defeats its purpose. Also, due to their sequential nature, BO algorithms are difficult to parallelise.

Evolutionary algorithms (EA) The last two algorithms covered in this section (genetic algorithm and particle swarm optimisation) are classified as evolutionary algorithms (EA). EAs belong to a set of modern heuristics based search method (metaheuristics) used in many optimisation problems [198]. The design of EA algorithms is inspired by the processes of evolution such as reproduction, mutation and natural selection. In general, EAs converges to a local minima, nevertheless may provide a sufficiently good solution while being able to explore high dimensional spaces at relatively low computational cost [29].

Genetic algorithms (GA) [112] are inspired by the biological processes where in each generation elements of the well-performing HP combinations are passed to the next generation. In the GA, every HPs candidate solution is represented by a chromosome which is encoded using string (genes). The common encoding techniques are binary, value and permutation strings [202]. The most popular encoding is binary as it has well defined operators (see below) while other encoding might require custom definitions. The initial population of chromosomes is randomly selected. The best performing chromosomes (*fittest individuals*) are than identified according to the value of the objective function. The chromosomes with better fitness are more likely to be passed to the new generation. Biologically inspired operations such as mutation (random modification to the binary string of genes), cross-over (partial exchange of genes between chromosomes) and inversion (partial binary string reversal operation) are used to generate new candidates. The process is repeated until the best subset is identified based on their performance. The convergence speed and accuracy of GA depends on the initial selection of the HP population, i.e., the initial random selection should include HPs which are close to the global optimum. The performance of GA, similarly to BO, is strongly correlated with the initial selection of its own HP such as crossover rate, mutation rate, initial population size and fitness evaluation function.

Particle swarm optimisation (POS) [102, 163] is another example of EA. In POS, a candidate solution to the optimisation problem is represented by a particle [214].

Every particle living in a HPs space is labelled and is described by the position vector $\mathbf{p}_i$, velocity $\mathbf{v}_i$ and its local best known position so far. The particle with the best global position informs others about the current optimal solution. Therefore the motion of particles is semi-random and governed not only by their local environment but also by the current best position of the entire population. The process is repeated until satisfactory solution is obtained or particles can move no longer (are stuck in one or more minima). As with any EA algorithm the performance is limited by proper swarm initialisation.

4.3 Methods

Motivated by the successful applications of GA for the discovery of efficient many-body potentials [86], we develop a two-stage fitting procedure where we attempt to converge to a global minimum of a custom made global loss function $\mathcal{L}_g$ in an iterative manner.

In our procedure, the model $\mathcal{M}$ architecture is varied automatically by the external optimisation algorithm subject to a custom set of physically motivated constraints. The composition of the training data base $\mathcal{T}$ is varied manually at this point in time. In the future, the aim is to introduce active learning algorithm to supplement the fitting procedure.

The set of physical constraints and the corresponding weights is selected by the user and represents prior knowledge about the system and the future intended use for the interatomic potential.

4.3.1 Global Loss Function

The functional form for a global loss (GLF) is given as

$$\mathcal{L}_g(\mathcal{M}, \mathcal{T}) = \sum_{\alpha} \omega_\alpha \mathcal{L}_\alpha(\mathcal{M}, \mathcal{T}) \tag{4.2}$$

where $\mathcal{L}_\alpha$ is a loss associated with the α constraint and w_α is a weighting parameter which represents its importance in the fitting procedure. For a

constraint loss function we choose the following form:

$$\mathcal{L}_\alpha(\mathcal{M}, \mathcal{T}) = \left|\mathcal{P}_\alpha(\mathcal{M}, \mathcal{T}) - t_\alpha\right|^N \tag{4.3}$$

Here, $\mathcal{P}_\alpha$ is a prediction on the α constraint with a true value t_α. The power N controls the type of a loss function, i.e., $N = 1$ results in an absolute loss, $N = 2$ gives commonly used quadratic loss function, and so on.

The weight factor w_α have inverse units of corresponding constraints α raised to N-power such that product in eq. 4.2 is unit-less. The interpretation of weighting parameters is intuitive, such that they control numerical precision of obtained loss for a given α relative to the remaining weights, i.e., doubling weight makes it twice as important as before.

4.3.2 Search Space and Performance Constraints

The global objective function does not only take into account the performance of the model with respect to the validation set but also allows to include constraints (labelled α in eq. 4.2) on the physical predictions of the model. Henceforth, the term *performance constraints* (PC) will refer to $\{\alpha\}$ as they directly control the predicting power of the interatomic potential. It follows that there are two types of PC which can be selected for the global loss function. First, those which are associated with the performance of the model on training or validation sets, i.e., energy, force and stress RMSE. The second group of constraints are those which are physically motivated. They represent desired properties of the model such as its ability to reproduce particular surface energy or the energy difference between two crystal structures.

On the other hand the *search space constraints* (SSC) term will be used to indicate constraints which are enforced on the model architecture more directly, such as number of Gaussians in a descriptor or a model's cutoff distance. In other words SCCs generate configurational space for HPs which will be explored by the optimisation algorithm to satisfy PC.

4.3.3 The global optimisation algorithm

The global optimisation algorithm (GOA) works in an iterative manner to optimise model architecture subject to training data, validation set, search space and performance constraints.

The optimisation procedure begins with the construction of the data sets. The current implementation of the algorithm assumes that the training and validation sets remain unchanged throughout the automated optimisation process as *Ta-dah!* does not currently provide tools to streamline this process. It is worth reemphasising that the training data set has disproportionate impact on a success of the potential. The main purpose of GOA is to optimise model's architecture, and in consequence improve its transferability. Next, target PCs are defined along with the SSCs in the configuration file. Every PC has weight associated with it which represents its relative importance in the fitting procedure. Once the initial selection is done, the iteration process consisting of three stages begins:

1. The global optimisation algorithm selects candidate HPs from a pool of SSCs.
2. The model is trained using new settings on the current data set.
3. Performance of the model is measured against PC.

The iteration procedure continues until predefined convergence criteria is achieved, such as a value of a GLF. Alternatively, the algorithm can be terminated manually or stopped after required number of executions. In all cases the best performing model at the time is available. The potential is therefore affected by changes in HPs.

The HPs selection process is controlled by *MaxLIPO+TR* algorithm from Dlib C++ library [103]. MaxLIPO+TR is a parameter free global optimisation algorithm which improves upon original LIPO algorithm [120]. The LIPO algorithm has been designed for optimisation of functions under the assumption that finite Lipschitz constant exists. A Lipschitz constant measures the maximum gradient of a function in a region over which the function is defined. It has been used previously to develop global optimisers which exploits surface smoothness and its regularity with respect to the input [140, 164].

The MaxLIPO+TR algorithm is capable of estimating Lipschitz constant which allows construction of a linear upper bound (upper as algorithm optimises to a global maxima) to the objective function. The algorithm selects point at random, evaluate its upper bound and compare it with the best current point. If the new point is better it is used in for the next evaluation.

The common issue with global optimisers is their slow convergence once close to the optima. The MaxLIPO+TR performance in resolving local optima is improved by implementing trust region method commonly used for derivative free optimisations [73, 171]. The trust region method fits a quadratic surface around the best candidate and then iterate this surface instead of the true on to quickly converge to a given region optima.

4.4 Results

To validate the global optimisation algorithm we employ it to a number of test cases. Even though the ML methods can, in principle, easily fit classical potentials, in practice by virtue of higher dimensionality it is only the case when sufficient training data are present to eliminate unnecessary features. The results presented here build upon on each other with increased complexity. The goal is to illustrate certain aspects of the global optimisation method and to establish its numerical limits. For the application of the fitting procedure to a more concrete, real-life scenarios the reader is referred to chapters 5 and 6.

4.4.1 (Re)discovery of a Lennard-Jones Potential

Perhaps the most basic, although not trivial, test is an attempt to "rediscover" a known potential. To do that the training data are generated using LAMMPS MD with a standard 12-6 LJ potential without a cutoff function (eq. 2.38). The parameters for LJ are $\epsilon = 0.043$ eV and $\sigma = 3.428$ Å which were fitted to model interaction between two argon atoms [94, 147]. The training data set consists of 11 snapshots obtained every 30 ps from the NPT simulation at T=10 K and atmospheric pressure after initial equilibration. The simulation box consists of 3x3x3 4 atom fcc cells and the cutoff distance is set to a generous 13.5 Å to reduce issues related to a lack of smoothness at the cutoff distance. The training data are suboptimal by design and clearly will not cover the full range of interactions even

for a simple two-body potential. Note that argon melting point is 84 K hence the training data consist of atoms oscillating around their equilibrium position.

The truncated LJ is chosen because LAMMPS' smooth versions differ in the implementation of the cutoff function as compared with *Ta-dah!* 's Mie-type descriptor which will be used for fitting. The Mie-type descriptor is composed of two components r^{-n} and r^{-m} and the dummy cutoff function is used to simulate truncated LJ.

4.4.1.1 The baseline

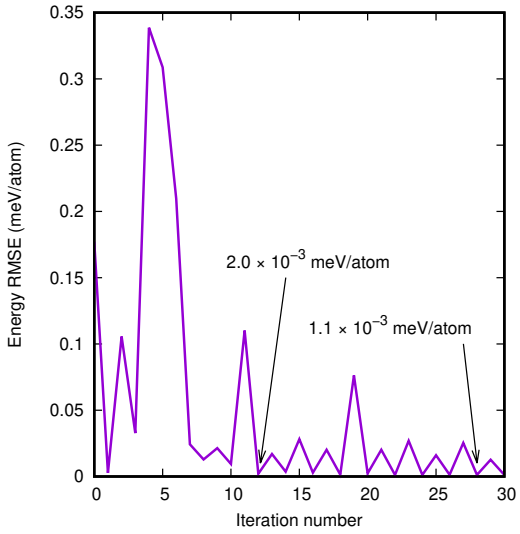
To establish the baseline the first test is performed to optimise GLF within a two dimensional HPs search space. The goal here is to obtain coefficients n and m such that the difference between predicted and true energies is minimised. It follows that GLF is simply equal to RMSE of energy. Intentionally, we do not restrict HPs m and n to integers but instead search a continuous [1.0, 20.0] range. The GOA almost immediately converges to expected result (fig. 4.1a).

At the 11th iteration the convergence to energies is below 2×10^{-3} meV/atom and after another 300 iterations the algorithm reaches its maximum accuracy of 1.5×10^{-9} meV/atom. The obtained Mie coefficients are $n = 11.9999989445$ and $m = 6.00000016379$.

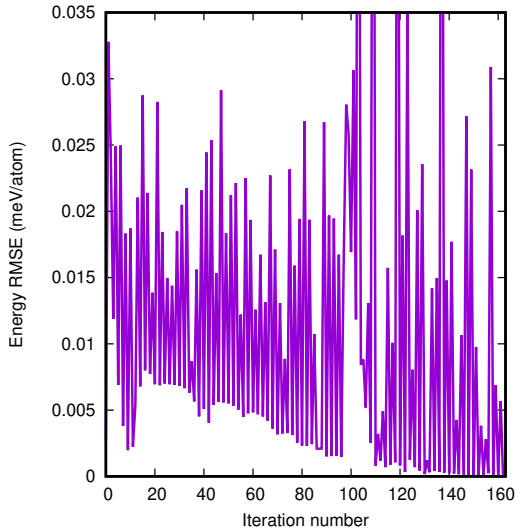
4.4.1.2 Optimisation with physical constraints

Here, GLF is not optimised against the RMSE in energies but against a lattice parameter $a_0 = 5.29173210729506$ Å and the cohesive energy $\epsilon_0 = -0.364688159928677$ eV. Note that, instead of analytically obtainable values for $a_0 = 5.28$ Å and $\epsilon_0 = -0.370$ eV we use values from the minimisation procedure using original LAMMPS LJ potential and keep maximum numerical precision for comparison purposes. The calculated target values differ slightly as compared with analytical ones because of the hard cutoff used. Note that the discrepancy almost disappears when long range interactions are accounted for with the increased cutoff distance.

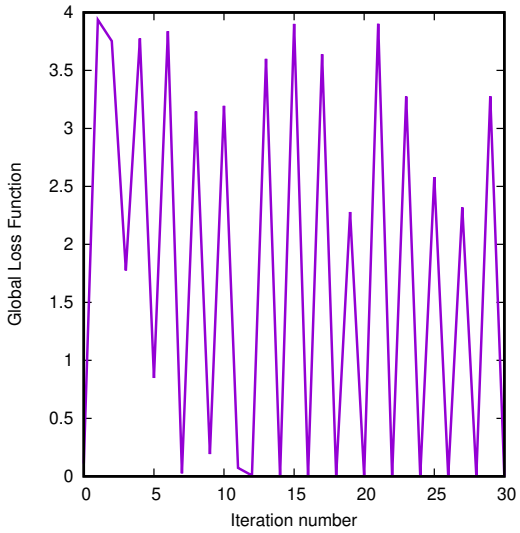
Similarly to the previous test, the convergence to the global minimum is achieved in around 12 steps with further iterations required to resolve HPs values to maximum numerical accuracy. The optimised coefficients are $n = 12.000122253$



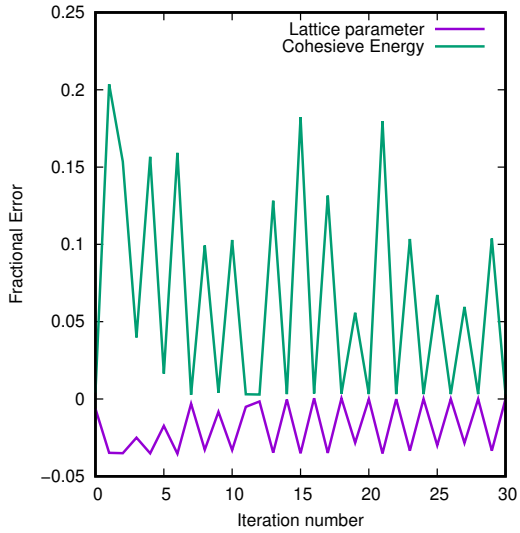
(a) Energy RMSE for the baseline test



(b) Energy RMSE for optimisation with PCs



(c) GLF for optimisation with PCs



(d) Fractional error in PCs

Figure 4.1: Convergence of the global optimisation algorithm for the two-dimensional hyper parameter search space. (a) Optimisation using energy only for the baseline test. The GLF which is being minimised is numerically equal to the energy RMSE. (b) Energy RMSE for optimisation with PCs. (c) For the optimisation with PCs, the GLF is a weighted squared sum of differences between predicted and true values for lattice parameter and cohesive energy. (d) The fractional error in lattice parameter and cohesive energy during the optimisation process with PCs.

and $m = 5.99994986975$. The predicted RMSE on energies is less than 3×10^{-7} meV/atom. The calculated values of the lattice parameter with obtained ML potential is $a_0 = 5.29173212504955$ Å and the cohesive energy is $\epsilon_0 = -0.364688207718215$ eV. The difference between predicted and target values is less than 2×10^{-8} Å and 5×10^{-8} eV respectively.

4.4.2 CCSDTQ Krypton

Having established numerical capabilities of the GOA in the previous section, the next step is to measure its performance against more sophisticated two-body function using mathematically more flexible ACSF (eq. 2.91) two-body descriptor.

The training data set is composed of calculations performed using the coupled cluster method with single, double, triple and quadruple excitations (CCSDTQ) and is available from [96]. The training data consists of bond energies for 36 separations between krypton atoms[2]. The CCSDTQ estimated energy uncertainty at the equilibrium distance is approximately 0.05 meV/atom and larger for smaller distances [96].

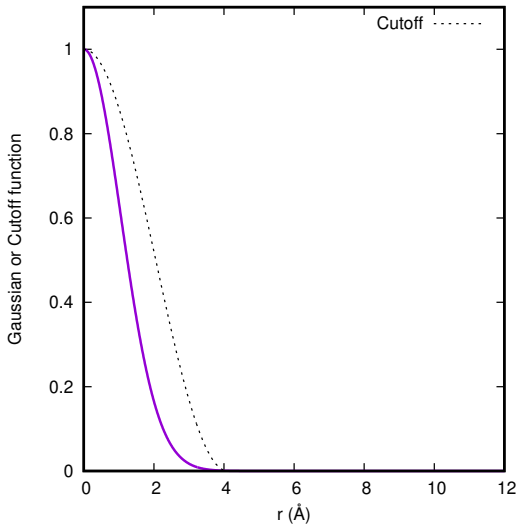
The ACSF two body descriptor is used along with a linear kernel resulting in a simple two-body model. Four separate MLIPs are developed with ACSF grids of one, two, three and finally four Gaussians. Therefore the fitting procedure is conducted separately for each MLIP and includes optimisation of each Gaussian position and width along with a cutoff distance r_{cut}. Every MLIP optimisation run begins with the same initial settings for the range of Gaussian positions ([0.0, 12.5] Å), widths ([0.001, 10.0] Å^{-2}) and cutoff distance ([4.0, 13.0] Å). As in the previous test those ranges are deliberately broad.

Grid	r_{cut} (Å)	δE (meV/atom)
1	4.1	4.9
2	13.0	0.91
3	13.0	0.089
4	11.5	0.011

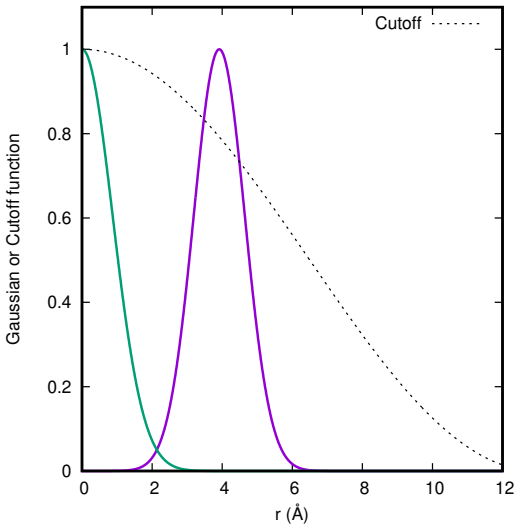
Table 4.1: Grids, optimised cutoffs and calculated errors on predicted energies as compared with CCSDTQ krypton data.

Figure 4.2 shows optimised ACSF grids for four different models. The corresponding potentials are presented in fig. 4.3. The optimised cutoff values and

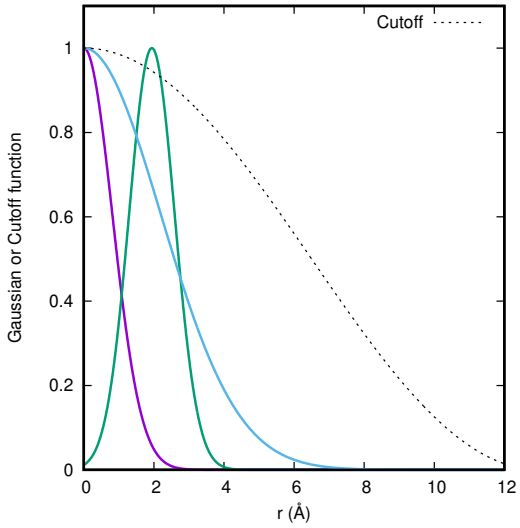
[2]The training database was prepared by fellow PhD student Asuka Nakamura-Pinder.



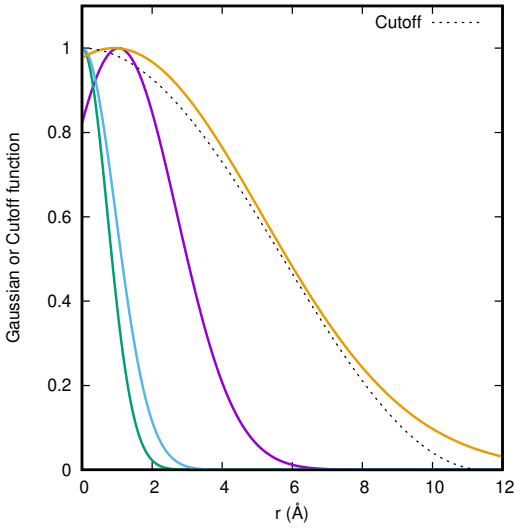
(a) Optimised grid of one Gaussian.



(b) Optimised grid of two Gaussians.

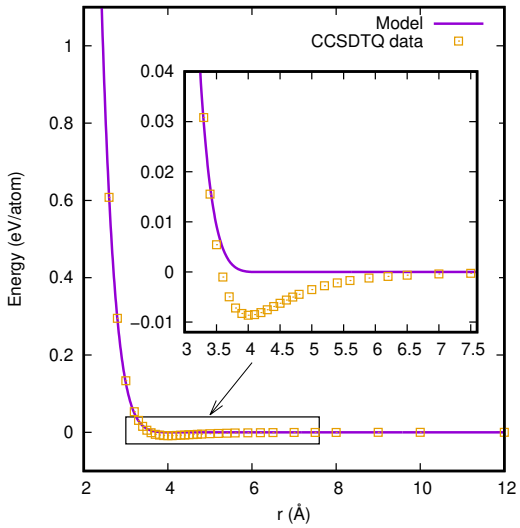


(c) Optimised grid of three Gaussians.

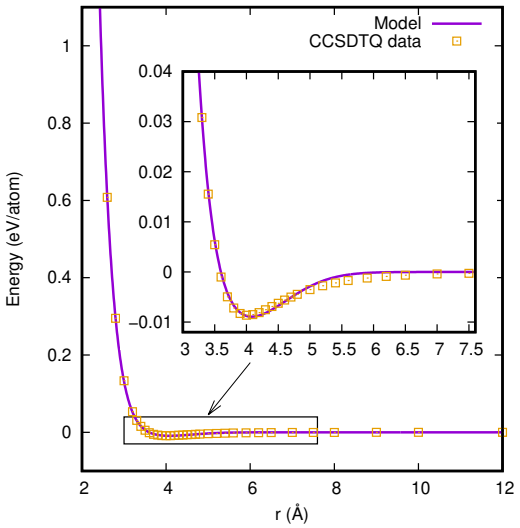


(d) Optimised grid of four Gaussians.

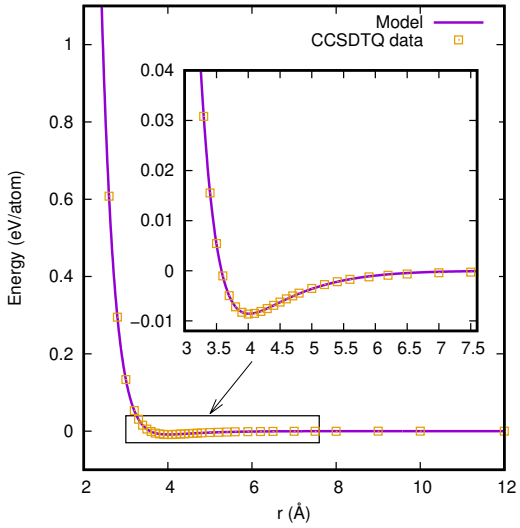
Figure 4.2: Optimised grids for ACSF descriptor for the CCSDTQ krypton data. For the corresponding ACSF MLIPs see fig. 4.3



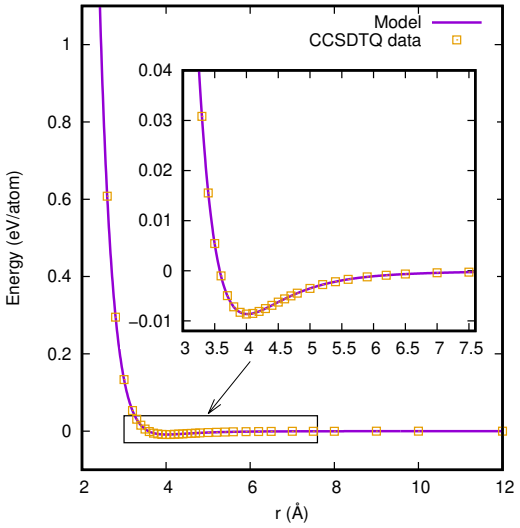
(a) MLIP with one ACSF Gaussian.



(b) MLIP with two ACSF Gaussians.



(c) MLIP with three ACSF Gaussians.



(d) MLIP with four ACSF Gaussians.

Figure 4.3: Two-body MLIPs obtained from CCSDTQ krypton data with global optimisation algorithm. For the corresponding ACSF grids see fig. 4.2.

the difference in the predicted energies of the models as compared with CCSDTQ data are tabulated in 4.1. In each case, the GOA successfully optimises ACSF grid to produce models with both increased complexity and better predicting power. The two Gaussian model already closely resemble original training data (4.3b). With just three Gaussians the obtained curve provides almost perfect match with respect to the training data while the model error of 0.089 meV/atom is close to CCSDTQ accuracy (4.3c).

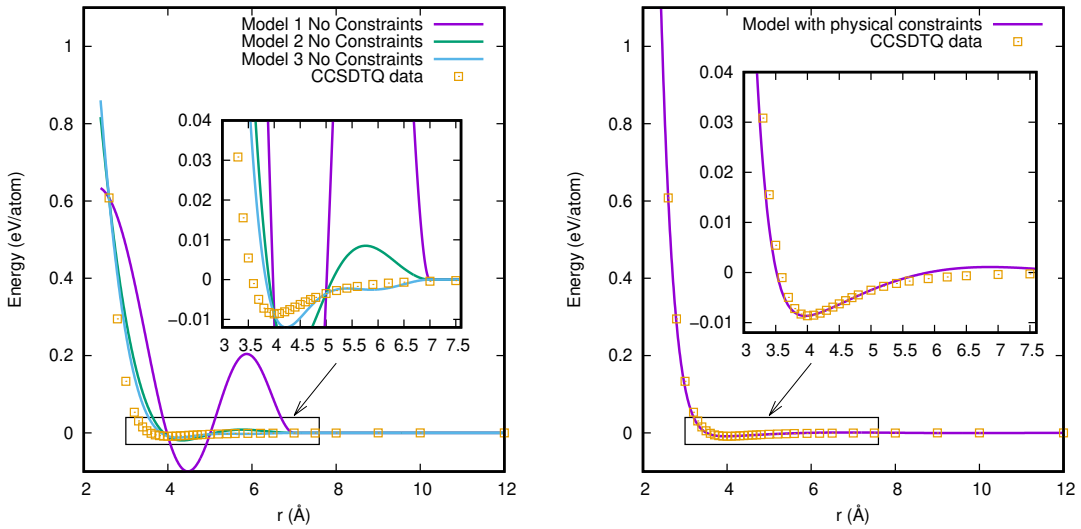
One more test is conducted to measure performance of the GOA when PCs are being used. To do that, a sparse training data set is constructed with just four data points selected from the original CCSDTQ calculations. The descriptor being used is a four Gaussian model described above.

In the first, deliberately naive, attempt, three different models are fitted simply by minimising energy RMSE. The result is shown in fig. 4.4a. The models clearly overfit data. All of them reproduce perfectly four data points, which were used during training, but fail catastrophically when faced with the unused CCSDTQ data. The fitting of each model begun with different initial settings. The best of the overfitted models used restricted HPs search space with very wide Gaussians while the other two used relatively broad ranges. Figure 4.4b shows a model which, in addition to fitting energy, RMSE is also being constrained by additional PCs. Here, PCs are the values of the lattice parameter and the cohesive energy calculated at different pressures (0 atom, 1 atm, 10 GPa, 100 GPa). The model shows clear improvement with respect to the models trained on energy RMSE only.

4.4.3 Many-body EAM

The last "relearning" case study is to evaluate the GOA on a many-body training data. To do that the training database is generated using general purpose EAM potential for tantalum [149]. The training data consist of three sets each containing 100 bcc configurations of 54 atoms sampled from the NVT MD simulation at 400 K (Ta melts above 3000 K) at the following normalised densities: (0.5, 1.0, 1.25). As in the previous cases the training data can be considered as insufficient to produce even a simple MLIP, not too mention a general purpose potential.

Two tests are attempted in this section. First is to relearn two-body part of



(a) Fitting with no constraints.

(b) Constraining function with PCs.

Figure 4.4: (Left) Figure shows intentional overfit of a relatively complex model with respect to the training data in the case where only energy RMSE is used during the optimisation procedure. The training data consists of just four data points which are always perfectly matched by three different models. However, the remaining, unseen, data are reproduced by chance only. (Right) The overfit problem is alleviated by introducing physical constraints (lattice parameter and cohesive energy) on the model.

the EAM model with ACSF two body descriptor. The second is the opposite of the first one, namely relearning many-body part with a custom modified EAD-like descriptor (mEAD) [183, 211] while the two-body part is computed by the original EAM model. The mEAD descriptor uses the $\eta\rho\log(\mu\rho)$ as an *embedding function* instead of the originally proposed square of the density, where η and μ are the fitting parameters. To clarify, in both tests the same training data are being used, i.e., there is no split into two-body data and many-body data.

This procedure is possible because *Ta-dah!* re-implements any EAM model two-body and embedding function as components of the descriptor. The functional forms are read from the *DYNAMO setfl* file in the usual way. This allows the fitting procedure to run with either one or both components, and mix it with any ML descriptor. For example, when both EAM descriptors are used and linear model is being trained on the data which have been generated with the same EAM model, then the re-learnt linear regression coefficients are *exactly* 1.0 for the two-body part and 1.0 for the many-body.

In both cases, the linear regression is against target energies and stresses and the regularisation parameter for the regression is estimated from the data by the

evidence approximation algorithm [26].

4.4.3.1 ML two-body descriptor

In here, two ML models are developed using ambient pressure data only (normalised density of 1.0). Both contain a many-body part from the original EAM potential. The two body part is modelled using ACSF descriptor with either three or four Gaussians. The GLF is proportional to the energy RMSE and only HPs controlling positions and widths of Gaussian functions are being optimised with the GOA. The obtained linear regression coefficients for the

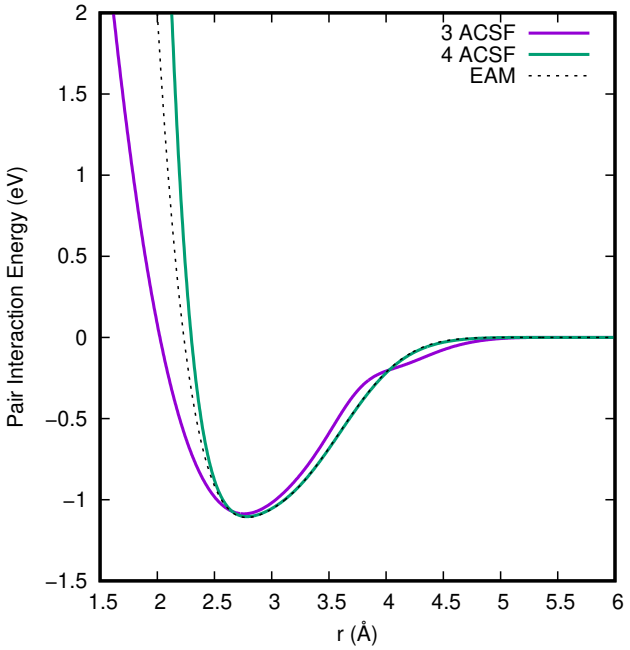


Figure 4.5: Pair interaction potential for ML models build with ACSF containing either three of four Gaussians and the original many-body function from the EAM model.

original embedding descriptor in models were 1.13 and 0.99 for three and four Gaussians respectively. The value close to 1 indicates good convergence of the machine learned two-body function given the configurations in the data set. Figure 4.5 shows relearned two-body part of the original EAM model and is compared to the original function. The model with four Gaussians shows excellent fit throughout apart from the repulsive section which is steeper as compared with the original pair interaction potential. This is unsurprising as only ambient pressure data are used during training, so this part of the potential is never used in generating the data. The simpler three Gaussian two-body potential reproduces the original curve only qualitatively.

4.4.3.2 ML many-body descriptor

Another four potentials are developed this time to fit many-body part of the EAM model. All models consist of the same initial architecture. The two-body part is represented by the original EAM pairwise function. The many body part is calculated with mEAD descriptor in the limit of the spherical charge distribution. The mEAD model contains five Gaussians each with the corresponding embedding function.

The embedding function varies with the density of the atomic environment therefore the first three models (ML no PC) are being trained using progressively more data [normalised densities $\bar{\rho}$: (0.5), (1.0, 1.25) and (0.5, 1.0, 1.25)] while the fourth model (ML PC) uses data generated only around the equilibrium density.

The first three models (ML no PC) developed use the same fitting process as with the two body function previously, namely to minimise energy RMSE. The fourth model (ML PC) attempts to compensate for the lack of spread in the density in the the training data. To do that, the potential is bounded by additional PCs such as lattice parameter, cohesive energy, unrelaxed vacancy formation and unrelaxed surface energy. The values used for constraints are obtained from the original EAM model. Therefore the minimisation of the GLF now takes into account not only model's ability to reproduce training energies and stresses but also loss associated with the difference between model's predictions and selected physical constraints.

The embedding functions obtained from all models are shown in figure 4.6 along the true one from the EAM potential. All models show excellent reproduction of the embedding function in the region close the the ambient condition density.

For models fitted without PC, addition of more training data is required to improve embedding function description in low and high density regions. The low density region is only qualitatively described and shows the expected energy gain when atom is embedded into the system however the exact shape of the functions differ as compared with the original EAM model. Addition of high density training data are necessary to obtain good description at high pressure. Inclusion of $\bar{\rho} = 1.25$ data shows qualitatively correct behaviour of an increasingly higher energy required to insert an atom into the existing atomic configuration.

The introduction of PCs improves the potential even further. The fourth model (ML PC) retains very good description of the embedding function at ambient

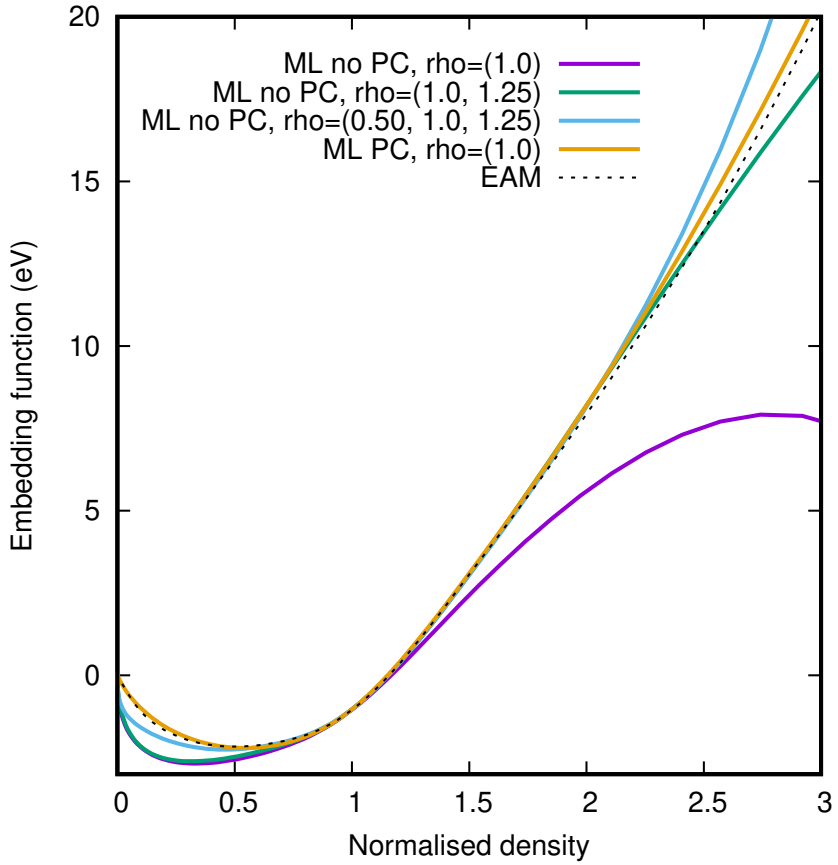


Figure 4.6: Figure shows the embedding function against the normalised density for bcc tantalum using Ta2 EAM potential from [149] (shown as the dashed line) and four different ML potentials. First three ML models developed without PC (ML no PC) are using progressively more training data at different densities. They show very good fit close to the training data but fail to extrapolate beyond known density range. The introduction of PC results in the model which resembles the original EAM models well across the full density range. Note that only density data around $\bar{\rho} = 1$ where used during the fitting procedure for the last model. See text for a description of the optimisation procedure for all models.

pressures and corrects low and high density regions appreciably despite the fact that it uses training data generated only at ambient pressure.

Note that the relatively small differences in the embedding function can have large effects on the model's performance. For example, the relaxed surface energies obtained with the best model without PCs are $2.07\,\mathrm{J/m^2}$, $1.75\,\mathrm{J/m^2}$ and $2.23\,\mathrm{J/m^2}$ for (100), (110) and (111) respectively. The constraint model improves significantly: $2.39\,\mathrm{J/m^2}$ $1.98\,\mathrm{J/m^2}$ and $2.63\,\mathrm{J/m^2}$. The first two values of the latter model are almost identical to the true values and the (111) energy is just $0.1\,\mathrm{J/m^2}$ lower [149].

4.5 Discussion

The fitting of the two-body function can be considered as an exercise in overfitting without the negative consequences, at least most of the time. Given enough data and relatively simple architecture of the two-body model, the naive optimisation of the HPs such that they minimise the energy difference between prediction and the training data is often sufficient as demonstrated in fig. 4.3.

The CCSDTQ data can be perfectly fitted with just four ACSF Gaussians while three Gaussians are already sufficient. The presented method is general and should work equally well for any two-body CC data. It compares well in terms of accuracy and the model complexity with the analytical potential function developed for krypton in [96]. The three ACSF Gaussian model is characterised by 10 parameters (as compared with 12 in [96]): every Gaussian is described by its position and width, there are three regression coefficients, and the cutoff function is parameterised with a cutoff distance only. Note that the parametrisation process of analytical potentials is often time consuming as it requires choosing the adequate model functional form. In contrast, the procedure presented in this work is accomplished within minutes while the optimisation process itself takes seconds for a simple two-body system (fig. 4.1c).

However when the data are sparse, optimising even the simple pairwise function shows negative effects of overfitting as shown in fig. 4.4a. The proposed solution of introducing PCs to compel the functional form of the potential and significantly alleviates the problem without a need for extra training data (fig. 4.4b).

One can see that the problem of choosing and then optimising model's architecture is inevitably coupled with the problem of picking the adequate training data sets. Simply put, complex models necessitate more data. This work attempts to tackle the optimisation of model's architecture given a fixed training data set. It is shown that the coupling is weakened by introducing external performance constraints (fig. 4.4b). Therefore, the optimisation with PCs leads to models with better transferability. Those two types of constraints complement each other during the fitting procedure. For example, sub optimal data set with a physically motivated model build out of relatively simple functions perform well. On the other hand, complex potentials require extensive data sets to constrain their functional form. Note that interesting physics often happens during the transition from one state to another. Those are rare events and constructing

representative data sets for the transition states is extremely challenging. In other words it is easy to wash out important features which affect the dynamics of the system (for example during the phase transition) with relatively unimportant equilibrium data. In general, PCs are experimentally or theoretically measurable quantities (macrostates). Therefore there is no need to attempt to come with a relevant set of microstates for a given phenomena.

The ability of the GOA to find the best performing set of HPs is further tested with EAM generated many-body data set. Once again GOA shows good ability to reproduce two-body function with four ACSF Gaussians (fig. 4.5). The discrepancy in the repulsive part can be attributed to the lack of data in this region as the training data were generated at 400 K and no PCs were used.The reproduction of the many-body part (in this case it is equivalent to the embedding function) is more challenging as very little variation in density is present in the training data set. Introduction of PCs is able to compensate for the lack of data as shown in fig. 4.6. Given scarcity of data the fit is very good and relearning the original embedding function is almost perfect as confirmed by the calculation of the relaxed surface energies.

It is shown that the GOA is capable of optimising HPs for complex functions. It is worth emphasising that its ability to do so can lead to overfitting when complex models are optimised to inadequate training data sets. This unwanted quality can be lessened by the introduction of PCs. Bounding by the physical constraints is just another way to control model complexity and improves its transferability.

The automatic optimisation of the HPs allows for the structured development of potentials and enable for the systematic comparison between different models. Moreover, it reduces the time required to optimise the model as compared with the manual trial and error tests.

The implementation of the GOA is not limited to a particular choice of the optimiser such as MaxLIPO+TR. Other viable options include particle swarm optimisation or genetic algorithms. However, the MaxLIPO+TR algorithm performs well because it allows for the degree of stochasticity in the objective function as well as having the ability to deal with function discontinuities.

The set of available physical constraints can be easily expanded and may involve more complicated computations such as the ability of the model to relax a vacancy. In principle, any properties which can be calculated by MD simulation can be used as PCs. There are a couple of practical limitations however. First,

the simulation must be computationally cheap as it will be performed at every iteration of the GOA. Second, occasionally very poor candidate potentials can give surprisingly good prediction on one PC while all the others are not so good. This false positive might force the algorithm into suboptimal local minima.

One has to balance the choice of PCs against available computational budget. Potentials with complex functional forms coupled with large data sets result in slow training and MD simulations. In consequence, each additional PCs increases computational effort to perform iteration of a global objective function optimiser. This can be partially countered by the bandit-inspired methods where the original training data set is split into smaller subsets which are then used for training and in consequence elimination of poor performing HPs. For example, optimisation of HPs for two body functions using energy RMSE is achieved within seconds on a desktop PC. Addition of a simple set of PCs, such as lattice parameter and unrelaxed vacancy extends this time to minutes. In case where the model utilises high dimensional descriptor and full set of PCs, the fitting procedure may lengthen to days or even weeks. In this case massively parallel supercomputers can be used, and may be essential, to drastically reduce those times.

It is also worth noting rather unwelcome oddity. In an attempt to automatically optimise model HPs we introduced different HPs in place of the original ones. Fortunately, this swap is not like-for-like. The initial HPs are often combined in the non-linear fashion hence difficult to interpret when manually tuning the model. The proposed HPs are linear weighting factors which reflect desired model properties and are simpler to interpret.

The choice of the loss function is somewhat arbitrary and based on our limited experience. We note that other choices are possible which can easily be implemented into our procedure. However the systematic assessment of those is beyond the scope of this work.

The two-stage fitting procedure takes advantage of a flexible functional form for the potential, capable of reproducing complex PES, and couple it with physical constraints to generate potentials with increased performance for configurations which are outside of the training set. The outcome of model fitting against energies, forces and stresses is highly dependent on the quality of the training data. The introduction of additional PCs enable important physical properties to be fitted correctly with far less data. This is because the PCs introduce very different, emergent properties which broaden the training beyond just forces of

atoms. It is possible for standard training sets to include forces from very different environment, but ultimately they are still only forces.

Chapter 5

Tracing the Frenkel Line in a supercritical molecular nitrogen

This work is conducted in collaboration with Dr Ciprian Pruteanu. Parts of this work are published in [143, 144].

5.1 Introduction

The fluid in the supercritical region was for a long time considered as a homogeneous state of matter. However, in 2012 it was proposed that it can be split into two distinct regions consisting of *rigid* and *non-rigid* liquid states [33]. The crossover line (or a narrow zone) where the "transition" occurs is called the Frenkel line (FL). It is postulated that the qualitative behaviour of the *rigid* supercritical state is *solid-like* while the latter is *gas-like*. It is worth emphasising that the FL does not represent thermodynamic transition but rather a dynamic one where a number of structural properties can change, such as the ability of the liquid to sustain shear waves, or the sudden change in the diffusion constant.

The FL can be considered as an extension of a long term debate about the nature of the liquid state in general. One approach based on the van der Waals model is to treat liquid state as an dense structure-less gas. The other approach, championed by Frenkel [65], is to consider liquid as a fast changing "crystal" where particles can oscillate for short periods of time when caged by the neighbouring particles while on a longer time scale they perform a random walk by jumping

between cages. Those jumps are fundamental to liquid ability to flow and their frequency are related to the liquid relaxation time τ as proposed by Frenkel.

It is worth noting that the FL differs to the concept of the Widom Lines (WL). First, the WL is defined as an extrema of a thermodynamic property (such as maxima in isobaric heat capacity [165]) past the critical point (CP). Second, all WLs terminate at the CP by the definition (they are the continuation of the discontinuities in thermodynamic properties on the phase boundary). Experimental evidence suggests that Frenkel Lines do not terminate at the critical point[145]. Third, different thermophysical quantities give, in general, different WLs while the FL crossover defines much narrower zone in the phase diagram.

Originally, it had been proposed that the FL can be identified by the sudden change in the diffusion coefficient [33]. This can be attributed to the reduction of the mean free path to the point where it is comparable to the particle size. In the classical limit the heat capacity of the liquid close to the melt curve is $3k_{\mathrm{B}}$ where $1k_{\mathrm{B}}$ comes from particles transverse excitations. Therefore it is predicted that upon crossing the FL the isochoric heat capacity should be reduced to $2k_{\mathrm{B}}$. Another criteria is based on the velocity autocorrelation function (VACF). In the *solid-like* region particles are caged by the neighbouring atoms and in result show oscillatory behaviour which can be identified from oscillations in VACF. In the free-flowing *gas-like* region VACF decays to zero monotonically without oscillation and the rate of it is proportional to the density.

The aforementioned quantities are readily available from computer simulations, such as molecular dynamics, but extremely difficult or even impossible to obtain reliably from the experiment. The FL crossover has also been experimentally defined and determined for molecules with a well defined Raman modes, such as methane [169] or systems with a single Raman active mode. However this criteria can be ambiguous in many cases and limited to molecular liquid. Since the FL transition leads to structural changes of the liquid it is advantageous to identify relevant markers which can be obtained experimentally. It has been shown that the diffraction measurements can be used to identify the FL crossover in the supercritical fluid in the static compression experiment [30]. Arguably, the most practical way to identify the FL is a saturation of the coordination number with increased pressure as demonstrated in the neutron diffraction experiments [142, 145].

The FL remains somewhat as a controversial concept. Lines on the phase diagram

normally represent thermodynamic transitions, or well defined turning points in a physical quantity (WL). FL does not. Still, the FL crossover leads to qualitative change in the system behaviour as well as its structure. Perhaps, it leads to a more fundamental question on how phase diagrams should be drawn.

Computational studies of the FL in the supercritical region would give a precise way to investigate whether there are theoretical features which can be used to define FL. However, they are limited by the lack of interatomic potentials capable of accurately reproducing potential energy surface of the interacting molecules. For example, the popular "optimised potential for liquid simulation" (OPLS) force field [100, 101] predicts solid rather than a liquid phase at 160 K at elevated pressures. This failure can be attributed to the fact that OPLS potential for nitrogen was optimised for ambient pressure. Moreover, OPLS N_2 is a Lennard-Jones two-centre potential with sites located at atomic centres. These models are, in general, inadequate to study systems with strong quadrupole interactions such as N_2. The results of using OPLS potential for high pressure critical liquid are published in [144].

We also implement five-centre site-site potential model (5CM) for nitrogen which was parametrised using accurate coupled cluster data [85]. However, our implementation of the model into LAMMPS is somewhat troublesome. The technical details and encountered model's limitations are discussed in section 5.2.

Given the shortcomings of the current models, in this section, a new machine learned interatomic potential for supercritical molecular nitrogen is developed based on high level quantum chemistry data and benchmarked against neutron diffraction experimental data. The new potential is then used to study N_2 at the range of temperature and pressure conditions to trace the FL line including its origin.

5.2 Five-centre site-site model for nitrogen

The model implemented in this section (5CM) has been developed by Hellmann to accurately describe the potential energy surface of two interacting N_2 molecules in the dilute-gas limit [85]. The N_2 molecules are represented as rigid rotors with a fixed bond length of 1.1014 Å. Each rotor consists of five sites. Those are

used to compute the interaction energy between two rotors by summing over all available 25 distances.

In practice, the model is represented by seven sites: five sites are massless and are used for the computation of the interaction energy. Two additional sites are reserved for nitrogen atoms separated by the fixed bond length such that the rotor has the moment of inertia equivalent to the N_2 molecule.

The functional form of the model from [85] is coded in Python such that LAMMPS [189] can utilise it using *pair_style python*. This solution proves to be extremely computationally inefficient. Fortunately, LAMMPS allows to generate tabulated potential from the Python model using *pair_write*. Such tabulated potential is then employed by the *pair_style table* which is many orders of magnitude faster.

There are however certain difficulties which are mostly associated with the fact that the original model has been fitted to molecular clusters. We observe that out of over 80 publications that have cited the original paper only one implements the five-centre model in a molecular dynamics setting [107]. Therein, the model is used to compute interactions for methane and nitrogen mixture at ambient conditions. Unfortunately, authors neither provide sufficient implementation details nor information of what molecular dynamics software they are using. We attribute this lack of usage in MD, of an otherwise popular model, to the following observations.

First, the original model was fitted to finite clusters, and had infinite range of interactions. Molecular dynamics simulations require models with finite range such that the interaction energy smoothly goes to zero at the cutoff distance. The usual way to make use of such a model is to implement it with a relatively large cutoff distance. This may work for a dilute gas, however, for condensed phase we are faced with additional difficulties. We find that even at 14 Å and the time step of 1 fs the sudden truncation of the potential energy function results in fluctuation of the total energy which leads to an unstable trajectory. We recognise that the trajectory can be stabilised by reducing the time step however this scales linearly with computational effort and we find convergence to be very slow. Besides, at this cutoff distance, the computation of nearest neighbour (NN) lists already results in the neighbour list overflow. Even though the maximum number of NN stored for every atom by LAMMPS can be increased from its default value, the resulting computational efficiency is drastically reduced. We find that increasing the cutoff length further did not sufficiently alleviate those

fluctuations. We conclude that the potential in the original form is unworkable for a high pressure fluid even when a long cutoff distance is being used.

We modify the potential by adding the cutoff function such that the computed forces at the cutoff distance are zero. The cutoff function is given by

$$f_c(r) = (r_c - r)^2 \frac{(r_c + 2r - 3r_i)}{(r_c - r_i)^3} \tag{5.1}$$

where r_i and r_c are inner and outer cutoff values respectively and r is the distances between two dimer sites. The cutoff function is constant up to r_i cutoff and only the last 1 Å is smoothly reduced to zero. Such a choice preserves the original form of the potential and only modifies its long distance interactions. We found that this reduces the fluctuations in the total energy to 0.02 eV/molecule with 1 fs timestep and allows us to cautiously use the model.

The 5CM is only used sparingly throughout this chapter. It is mainly utilised as an additional verification tool for our new machine learning model which is developed in the following chapter. We also anticipate that the 5CM model is to no purpose to any free energy calculations.

5.3 Development of N_2 interatomic potential

The standard Kohn-Sham density functional theory (DFT) is the usual workhorse for the generation of training data sets during the development of the machine learned potentials. However, it is well known that the vanilla DFT does not provide accurate description of weak dispersion forces and requires specialised dispersion correction schemes which are often questionable in case of their transferability between different systems.

The aforementioned issue is magnified for a relatively simple element as nitrogen where the energy of the N-N triple bond is orders of magnitude higher than the weak dispersion forces between interacting molecules. At the same time it is recognised that the quantum chemistry methods, such as Coupled Cluster (CC), can provide extremely accurate potential energy surfaces (PES) albeit only for small systems, such as two interacting molecules, at very high computational cost.

Here, the new interatomic potential for nitrogen is developed based on the quantum chemistry data obtained from [85]. The potential is specifically designed

to reproduce available neutron diffraction experimental data for the supercritical N_2 at 160 K and 300 K [142, 145]. The potential is then employed to trace the FL in the P, T phase diagram.

While the underlying training data are four-body in nature (two interacting N_2 molecules with fixed bond length) the resulting potential is technically many-body. Such a choice allows to use the *Ta-dah!* universal LAMMPS interface without any modifications. There is however a trade-off as the many-body potential which will be used is density driven but the CC data do not provide this information. Therefore, to bypass this shortcoming, a bespoke training procedure is developed.

It is also worth noting that currently there is a lack of *ab initio* data for interactions between three nitrogen molecules. While in principle it is possible to include classical correction for the van der Waals forces (for example from the Axilrod-Teller-Muto three-body potential [8]) it is not strictly necessary. First, it has been shown that possibly due to fortuitous cancellation of errors the two-body (molecule-molecule) potentials can perform remarkable well when compared with the available experimental data [110]. Second, it is plausible to expect three-body interactions to be extremely faint as compared with already weak two-body dispersion forces. Third, the inclusion of the three-body term will increase significantly the computational effort of the simulation.

5.3.1 Training database

The training database is built upon publicly available quantum chemistry data [85]. The CC method with single, double and perturbative triple (CCSD(T)) excitations was used to compute four-dimensional PES of two interacting N_2 molecules. Further CC higher-order corrections were included as well as core-valence and relativistic effects (see [85] for detail). The data set consists of 408 data points with 26 different angular configurations and a range of distances. The bond length was fixed at 1.1014 Å. Furthermore, a five-site analytical model is provided in the paper and fitted to CC data which gives excellent fit. However, we do not utilise this model directly but test *Ta-dah!* ability to reproduce experimental data when trained on QM data as generated by the CC model. First, this allow us to benchmark *Ta-dah!* ability to reproduce models where energy differences between different molecular orientations smaller than typical error in DFT calculations. Second, by adjusting hyper-parameters we are able to

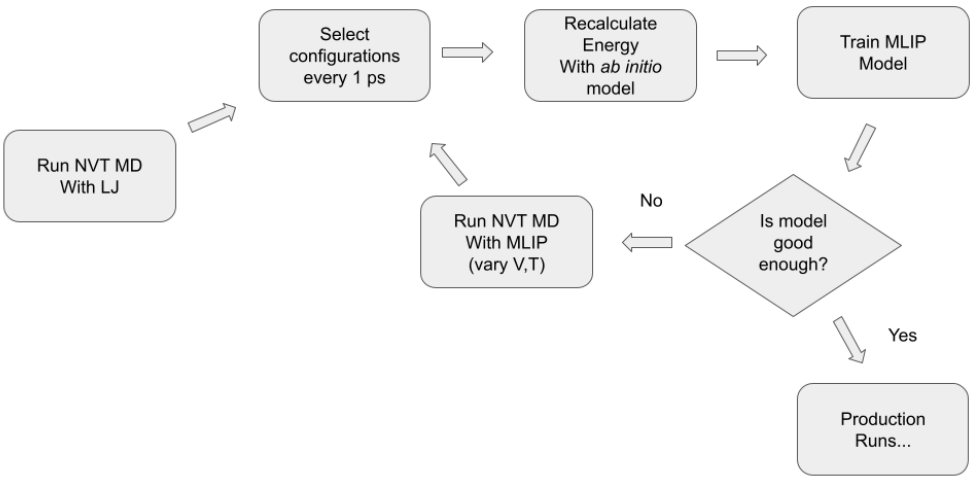


Figure 5.1: The iterative process of building machine learning interatomic potential for nitrogen. The optimisation process ceases when model satisfactory reproduces the NIST equation of state [172] and the experimental pair correlation function [142, 145].

fine-tune our model to reproduce experimental data. Third, we are able to train our model with a very smooth energy cutoff alleviating relatively high energy fluctuations encountered in the 5CM model (section 5.2).

As indicated in the previous section the development of the training database is crucial to satisfy the many-body nature of the descriptor (see the following section for model architecture). Here, instead of using CC data directly, the analytical model [85] is used to compute energies for a range of configurations. The development of the training database follows an iterative process as illustrated in fig. 5.1. In the first iteration of the fit, the initial set of reference configurations is obtained by sampling MD trajectories at the range of pressures and temperatures. The temperature range used to generate trajectories vary between 100 - 1300 K while the simulation box length varies between 6.1 and 15 Å. Note that the simulation box does not consist of solid phases and only liquid and gas data are used. Molecule-molecule interactions are computed using classical two-site 12-6 Lennard-Jones potential with the N-N rigid bond being fixed at 1.1014 Å. The NVT trajectories are sampled every ps to construct initial set of training data. In the next step the LJ energies and forces of the selected configurations are discarded and energies are recalculated using five-site analytical potential. The obtained data set is then used to train a first iteration of the model. In the following iterations, the machine learned model is used to generate more

MD NVT trajectories which are again sampled and energies are recomputed with an analytical potential. The model's architecture is also being fine-tuned at this stage. The performance of the model is monitored against the NIST reference equation of state [172] and pair-correlation functions obtained from neutron diffraction experiments [142, 145]. Perhaps unsurprisingly, the addition of new training data for a given P, T conditions improve the performance in this particular region. However it was found that if too much emphasis is given to a narrow P, T space the model's ability in other regions may suffer. This is likely to be due to the relatively simple model's architecture which is described in section 5.3.2.

The procedure is repeated until satisfactory fit is obtained. Note that this training data is chosen because the main motivation behind the development of this potential is to study the structure of supercritical nitrogen.

5.3.2 Descriptor and regression choice

The local atomic environment is captured by a combination of two-body (eq. 5.2) and many-body descriptors (eq. 5.4) as implemented into *Ta-dah!* package. The total energy of the system is simply obtained by summing individual atomic energy contributions. The maximum interaction distance between two atoms is set at 10 Å.

The pairwise interactions are expanded using blip basis functions B_n (eq. 3.5) while the cosine function f_c (eq. 2.92) ensures smooth energy cutoff

$$v_n(r_{ij}) = B_n(r_{ij}) f_c(r_{ij}) \tag{5.2}$$

where r_{ij} is the distance between atoms i and j.

The many-body interactions are then captured by first computing local atomic densities using Gaussian Type Orbitals [183, 210] (eq. 5.3):

$$\psi^{\eta,r_s}_{l_x,l_y,l_z}(r_{ij}) = x^{l_x} y^{l_y} z^{l_z} \exp\big(-\eta|r_{ij} - r_s|^2\big) \tag{5.3}$$

where η controls the width of the Gaussian function and r_s its position on a grid. x, y and z are components of the displacement vector between two interacting atoms. Exponents l_x, l_y and l_z are the quantised directional-dependent angular momenta. The sum of those define the angular momentum $L = l_x + l_y + l_z$ in

eq. 5.4 which effectively determine the order of the expansion. The summation is constrained as below (eq. 5.4 to ensure rotational invariance of the descriptor

$$\phi^{(i)}_{L,\eta,r_s} = \sum_{l_x,l_y,l_z}^{L} \frac{L!}{l_x!l_y!l_z!}\Big(\sum_j \psi^{\eta,r_s}_{l_x,l_y,l_z}(r_{ij})\Big)^2 \tag{5.4}$$

In practice, this many-body expansion is truncated at $L = 1$. In principle this results in a three-body descriptor (the expansion up to the p-orbital) which does not uniquely describe i-th atom local atomic environment. However, it is found sufficient when combined with a linear regression with a second order polynomial basis functions. By taking the combinations of its components, an accurate description of the four dimensional PES of two interacting N_2 molecules is obtained. The Bayesian approach to linear regression is used to optimise model coefficients along the evidence approximation algorithm [26] to control the regularisation parameter. This automated procedure of controlling model complexity ensures the model's transferability given sufficiently large training database. The Bayesian probabilistic models are generally prone to be overconfident in their ability to predict future data. Given that the generation of the database is computationally economical and covers wide range of P, T conditions, the evidence approximation algorithm seems to be working well.

5.3.3 Comparison with experimental data

The FL is related to dynamics of the system and its structure, which are not directly measured experimentally. Therefore it makes sense to gauge the model's performance against available experimental data. The interatomic potential's ability to model accurate radial distribution functions (RDF) can be compared to the neutron diffraction experiments performed by other members of the research team [144] The measured quantity in the neutron diffraction experiment is a structure factor S($\mathbf{q}$). The structure factor describes how incident radiation is being scattered by the atom. The wavevector transfer $\mathbf{q}$ represents the difference between the scattered and incident beam wavevectors. The experimental RDFs are then obtained by post-processing raw S($\mathbf{q}$) data. The coordination numbers are then readily available by either integrating RDFs to a fixed cutoff distance or the local minima after the first peak. The latter criterion is more suitable than a fixed distance because varying densities are being considered, and is used in both experiment and this work.

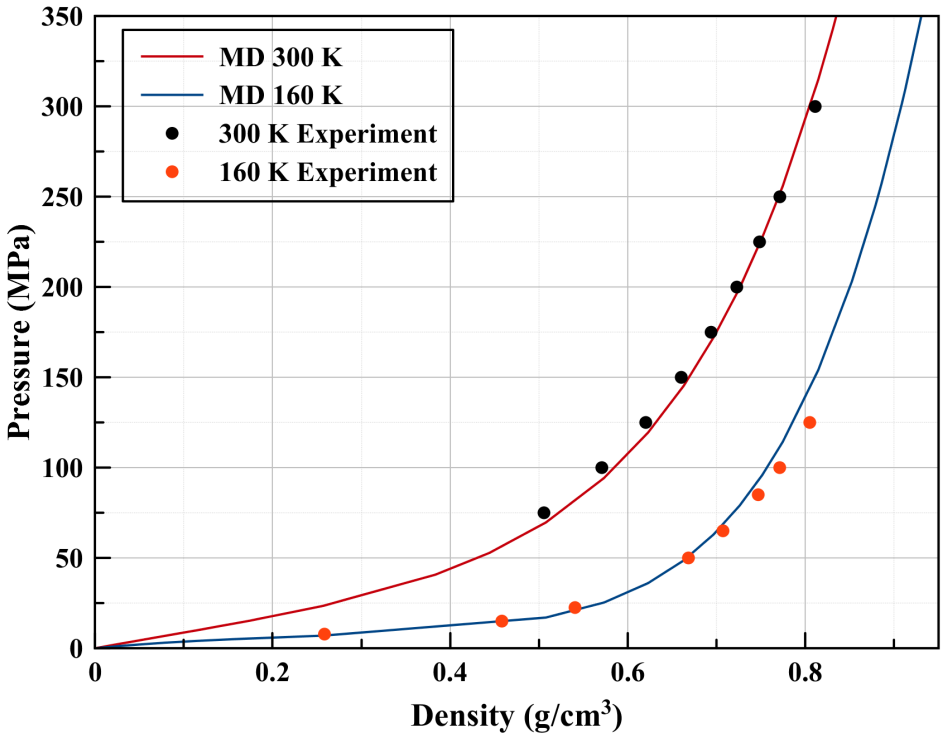


Figure 5.2: Equation of state for newly-developed machine learning potential compared with NIST equation of state reference [172] at the pressures where the experimental RDFs were measured.

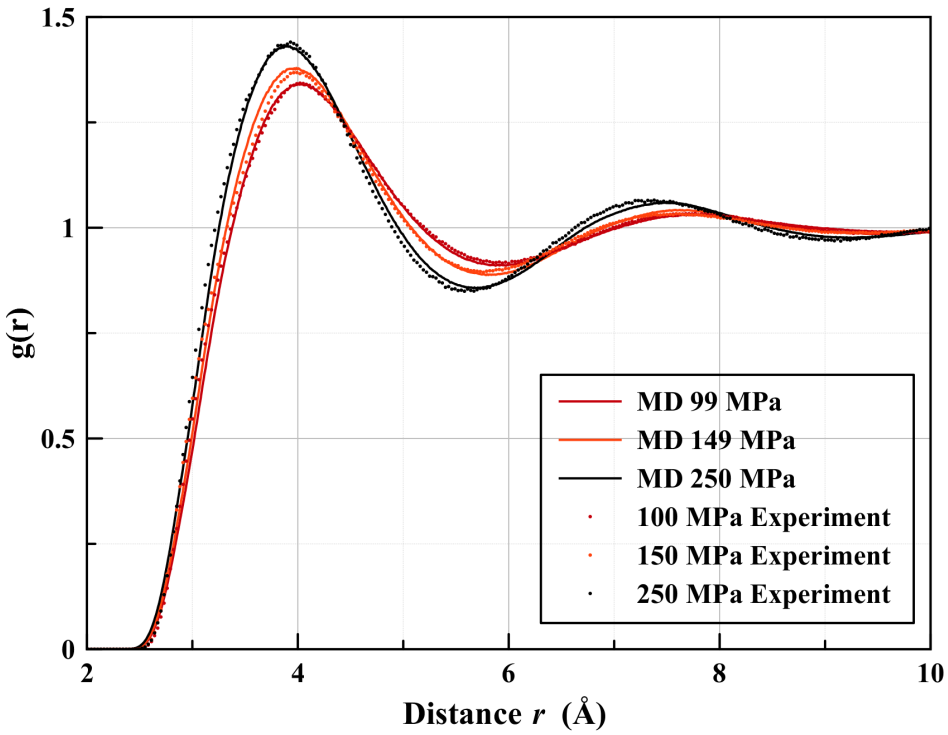


Figure 5.3: A comparison of experimental and MD radial distribution functions for supercritical nitrogen at 300 K using the ML potential.

Fig. 5.2 shows comparison of isothermal equation of state (EoS) obtained from MD simulations using LAMMPS package [189] with National Institute of Standards and Technology (NIST) reference [172]. NIST EoS was used for calibration of the neutron diffraction experiments at 160 K [142] and 300 K [145]. This experimental data are also used in this work.

In general, the simulated EoS is in good agreement with the NIST reference data. However there are some small discrepancies worth pointing out. At 300 K and low densities the model predicts lower pressures while at the high densities it is the opposite. The 160 K isotherm agrees well until 0.7 g/cm^3 but then progressively overestimate density for a given pressure. The reported differences are likely to be caused by the relatively simple model's architecture rather than insufficient training data, as those included configurations with densities over 1 g/cm^3. We do not attempt to train another model with a more complex architecture as this will increase the computational effort during MD simulations. We deemed the current model to be sufficient for the task at hand.

Comparison of experimentally obtained and simulated RDFs (fig. 5.3) shows

very good agreement. In particular the onset of g(r), the height of the first peak and the following minima are all matching the experimental data. There seems to be a general trend for the model to report first minima at larger distances in particular for higher pressures. This is consistent with previously observed small discrepancy between model and NIST EoS. The remaining RDFs are available in Appendix A.1. The small discrepancy of the model as compared with the experimental RDFs could be due to the limitations of the underlying QM data even though the model's HPs are optimised to reproduce experimental data.

Fig. 5.4 shows comparison of the coordination number between experiment and model plotted against both pressure (left fig.) and density (right). The MD data are smooth owing to sufficiently long production runs (2 ns) and simulation boxes containing over 2000 molecules. Even though the experimental data shows fair amount of noise, which is normal for neutron diffraction experiments, one can observe qualitative agreement between both curves. While the experimental data shows sudden flattening of the coordination number around $0.7\,\mathrm{g/cm^3}$ for both isotherms the simulation data indicates much smoother transition. This disagreement between two curves is likely to be due to differences between model and experimental RDFs as well as a small but noticeable discrepancy between NIST and model equations of state.

The change of slope of the coordination number as a function of pressure was used previously to identify the FL crossover and the same criterion is used in this work. Since the model shows good agreement with NIST EoS, experimental RDFs and, in consequence, coordination numbers, it is therefore more than sufficient to study FL.

The coordination number for crystalline solid is defined as the number of first nearest neighbours of an atom. Therefore, regular hexagonal close packing would yield maximum coordination number of 12. For a liquid the first coordination number is obtained by integrating radial density function from zero to the first minima which may result in a higher coordination number. This phenomenon exists even in simple discontinuous potentials such as as hard sphere model and the square well potential [143].

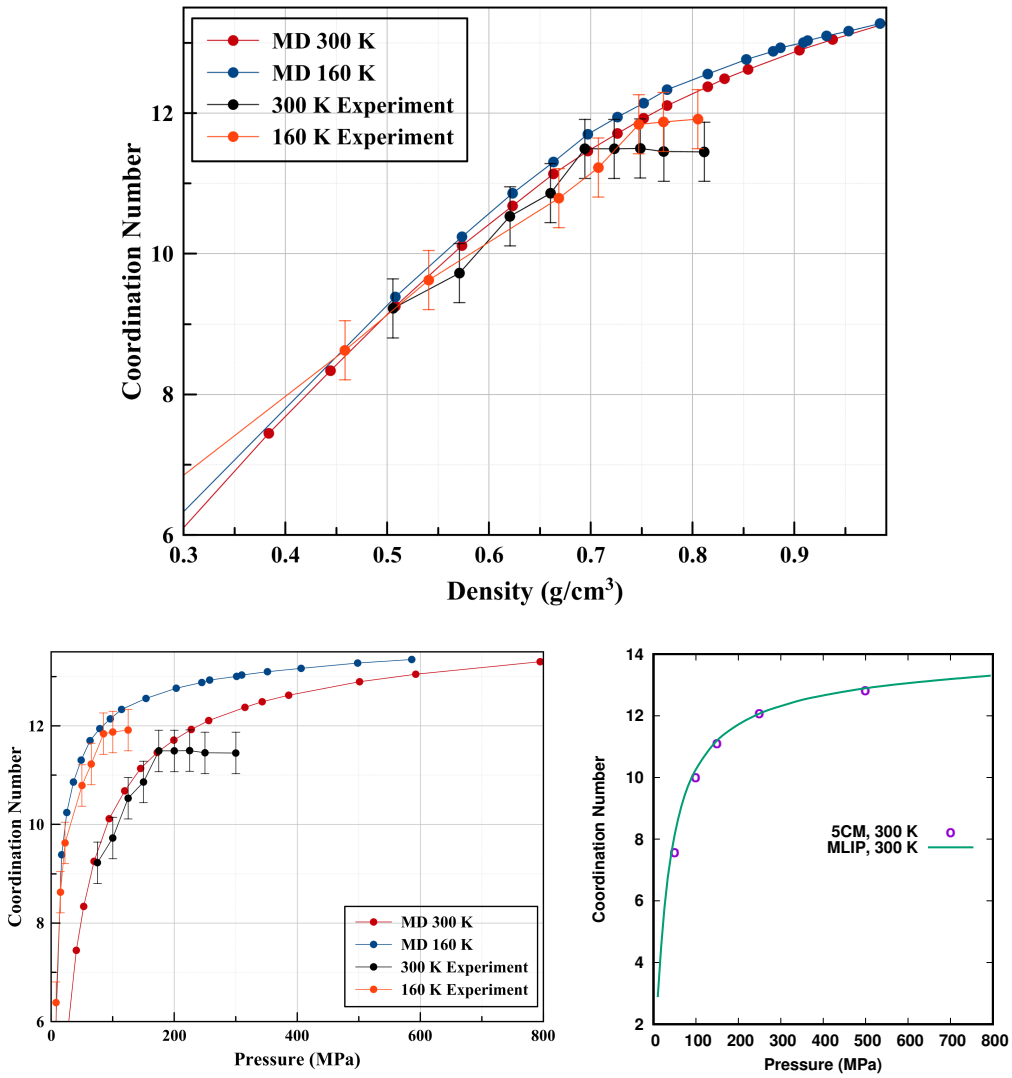


Figure 5.4: (Top) Coordination number as a function of density obtained for MLIP model compared with neutron diffraction experiment. (Bottom left) Coordination numbers as a function of pressure. The liquid phase can have the coordination number greater than 12 (see text for details). (Bottom right) Figure compares CN as obtained using two different models at 300 K. MLIP data are presented as a solid line, 5CM data are drawn as circles. There is an excellent agreement between models.

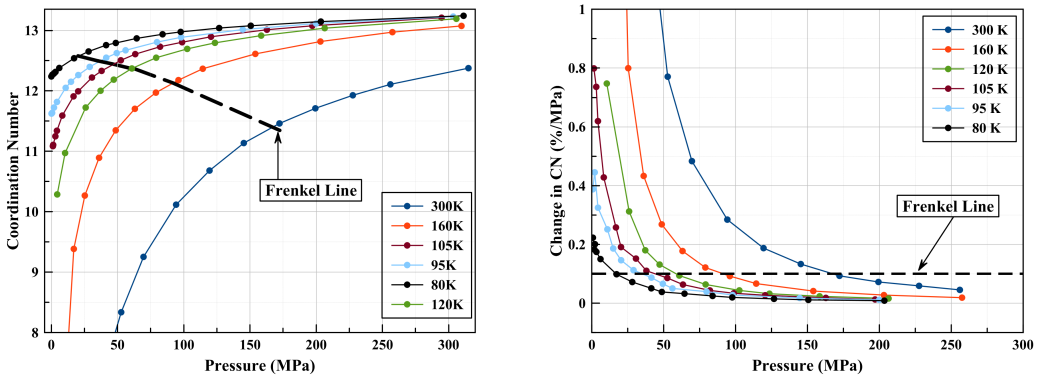


Figure 5.5: (Left) Coordination number of nitrogen from ML potential MD as a function of pressure. (Right) Percentage change in the coordination number of N_2 with increasing pressure.

5.4 Nitrogen Frenkel Line Terminates at the Triple Point[1]

The existence of the FL has been demonstrated both theoretically and experimentally. However, its exact position in relation to the critical point on the P-T diagram and in consequence its relation to the Widom lines was still unknown prior to this work. This is partially because most experimental studies have focused on the supercritical fluid and neglected the subcritical region. There is still an ongoing debate as to whether the FL is a different phenomenon as compared to the maxima/minima of thermodynamic quantities represented by Widom lines [30, 165]. Since Widom lines originate at the CP, the identification of the origin of the FL could unambiguously resolve this point of contention.

The identification of the exact point where the FL crossover occurs for a given P, T conditions is rather problematic as it is a dynamic continuous transition between *gas-like* and *solid-like* liquid structures rather than sharp thermodynamic jump. Two related quantities are therefore monitored as a functions of pressure, that is the evolution of the coordination number as well as the diffusion constant.

The coordination number is obtained in the same way as in experiments [142, 145], that is by counting atoms up to the first nearest neighbour shell. This is computationally achieved by integrating RDFs up to the minima which follows the first peak. The diffusion constant is obtained from the mean square displacement

[1]I performed MD simulations and partial data analysis, Dr Pruteanu carried out final data analysis including identification of the origin of the FL at the triple point.

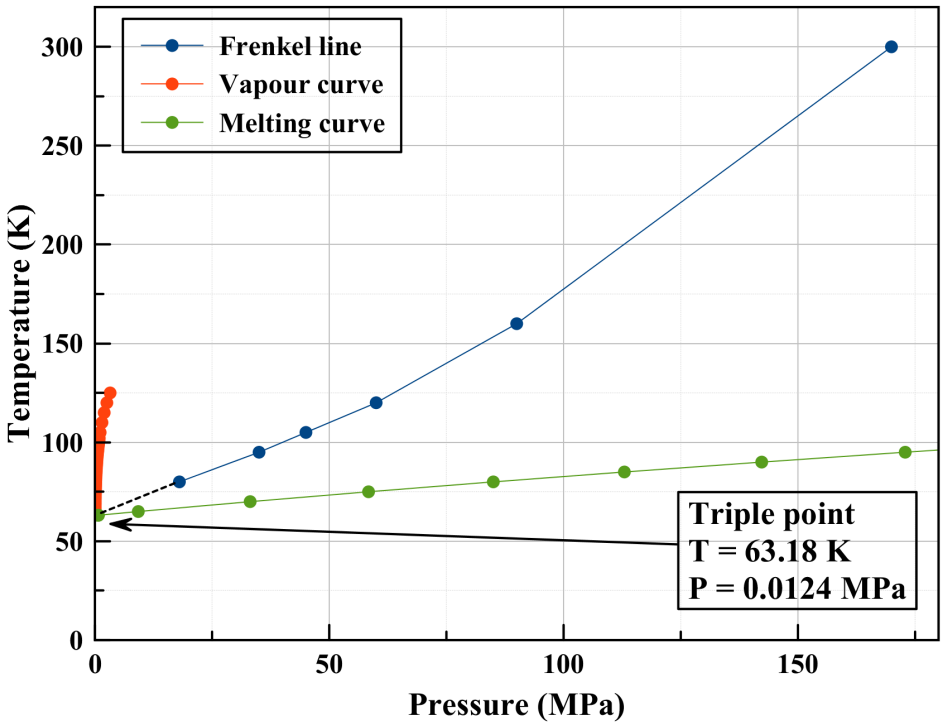


Figure 5.6: Current line as identified using correlated changes in the coordination number and the diffusion constant of nitrogen. The vapour and melting curves are depicted according to data from NIST [95]. The dashed line is a guide to the eye.

using MD trajectories. There is no diffusion experimental data available at the time.

FL is identified by the plateau of the coordination number or the change in the diffusion constant as the pressure increases. The evolution of the diffusion coefficient is shown in fig. 5.7 (left). As the pressure increases, diffusion constant abruptly drops. After the initial drop its value stabilises and the decrease in its value is much smaller with pressure.

The diffusion vs pressure curves are considerably less smooth as compared with the coordination number data, in particular at the low temperatures. The noise at low temperatures makes it difficult to provide analytical criterion as changes in the diffusion constant are much more subtle. The region where the rate of change of the diffusion constant decreases abruptly and then stabilises is theorised to coincide with the FL.

The coordination number is shown in fig. 5.5 (left). The same trend is observed as with the diffusion constant, namely the saturation of the coordination number with pressure. However precise identification of the FL position is somewhat more challenging as it is not a thermodynamic phase transition but a continuous crossover. To date there is no rigorous criterion available. Therefore to better quantify at what pressure the crossover happens the following, somewhat subjective, metric is proposed

$$\frac{P_{TP}}{C_N} \times \frac{dC_N}{dP} < 10^{-5} \tag{5.5}$$

where C_N is the coordination number and P_{TP} is the triple point pressure. Another way to identify the FL region is a straight line at very low densities, and a straight line coming through the coordination numbers at the highest densities. In the middle the lines will not fit the data, and that middle region is where the change is happening. Our criterion simply quantify this region such that obtained number is always, consistently, within this region.

The aforementioned criteria are applied to identify positions of the FL for a set of isotherms between 80 and 300 K. As expected, both criteria correlates well as shown in fig. 5.7 (right). Computationally obtained FL as a function of pressure is plotted together with experimental vapour and melting lines [95] and presented in fig. 5.6. The data indicate that the FL terminates at the triple point. Given that the Widom lines originate from the critical point our data suggests that

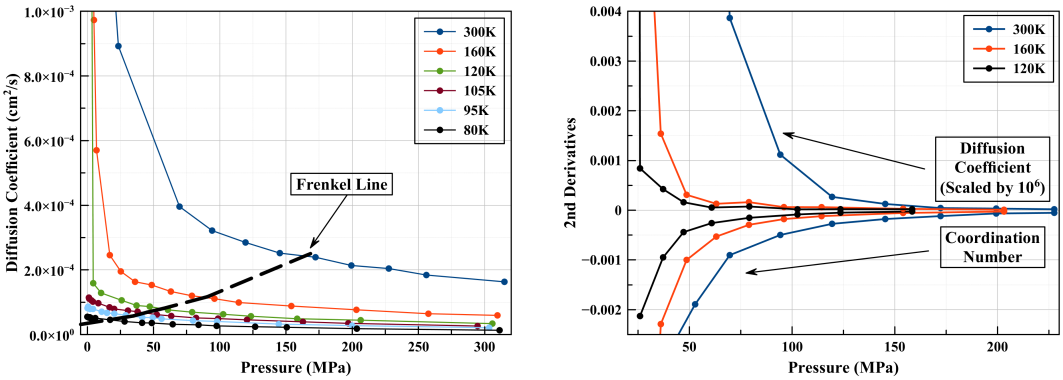


Figure 5.7: (Left) Diffusion coefficient from ML potential MD as a function of pressure and Frenkel line as determined from the coordination number equation below. (Right) Combined second derivatives of diffusion coefficient (scaled by 10^6) and coordination numbers as functions of pressure, at 120, 160 and 300 K.

the FL in a separable phenomenon to the Widom line. Therefore along any subcritical isotherm, the gas initially condenses to a non-close-packed liquid and only subsequently to a rigid liquid with medium-range order [144].

Chapter 6

Development of an improved interatomic potential for N_2. Application to the phase diagram.

The pressure-temperature phase diagram of nitrogen is extremely complex. On one hand, the triple bond of nitrogen is very strong with a dissociation energy of 9.72 eV/molecule [176]. On the other, the free energy differences between competing crystal structures can be extremely small, in the meV/molecule range. The molecular stability of nitrogen persists to over 100 GPa where the molecular bond dissociates and either an amorphous solid [58, 74] or the crystalline cubic gauche structure [57] is observed to form depending on exact P, T conditions. Between ambient conditions and 10 GPa, six solid phases have been confirmed experimentally.

This chapter is structured as follows. The model developed in section 5.3 is deployed to study the melting curve of nitrogen. It is found that the model is insufficient to study the phase diagram of nitrogen. The new model is then developed in section 6.2 to address the shortcoming of the first model. Next, the relevant solid phases of nitrogen up to 10 GPa are reviewed in section 6.3. Finally, the improved model is utilised to study the phase diagram of nitrogen in section 6.4.

6.1 Nitrogen Melting Curve[1]

The computation of the melting curve with an interatomic potential (and the phase diagram in general) provides an excellent benchmark for model capabilities. It aids prospective model users when deciding whether the model is suitable for intended computation. Arguably, for MLIP, it is of even higher importance as compared with classical interatomic potentials given the former known issues with poor extrapolation ability. It is a particularly stringent test of the transferability of the current potential because no crystal data was used in the fitting.

Given encouraging agreement between model developed in chapter 5 and experimental data in the supercritical region and its novel prediction of the FL termination at the triple point, the model is further deployed to obtain the melting line of nitrogen as a function of pressure.

Molecular dynamics simulation using the Z-method [20, 21] are performed to obtain the low pressure nitrogen melting curve. The obtained data are compared with the experiment. In the Z-method, a number of MD simulations are performed in NVE ensemble. Each simulation has incrementally increased internal energy and begins with the initial perfect lattice configuration. The low energy systems will stay solid while above the melting line the solid will be initially superheated and then melt given sufficient run time. The lowest temperature of the melt is then taken as a proxy of the *true* melting temperature. An example from this work is shown in fig. 6.1. The advantages of the Z-method include simple running process and the simulation cells with a relatively small number of atoms.

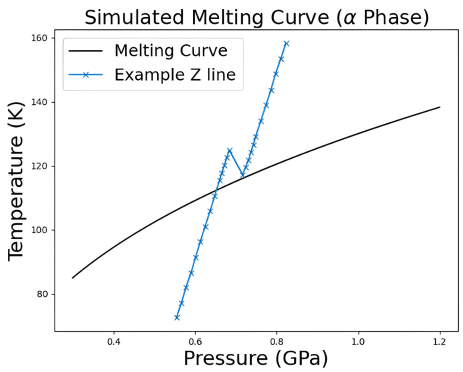


Figure 6.1: Illustration of the distinctive Z-curve in the Z-method. The melting temperature is taken as the lowest temperature of the liquid.

Up to approximately 10 GPa the relevant initial crystal configuration for the

[1]The computation of the melting curve was performed by Elspeth Smith as part of her MPhys project report.

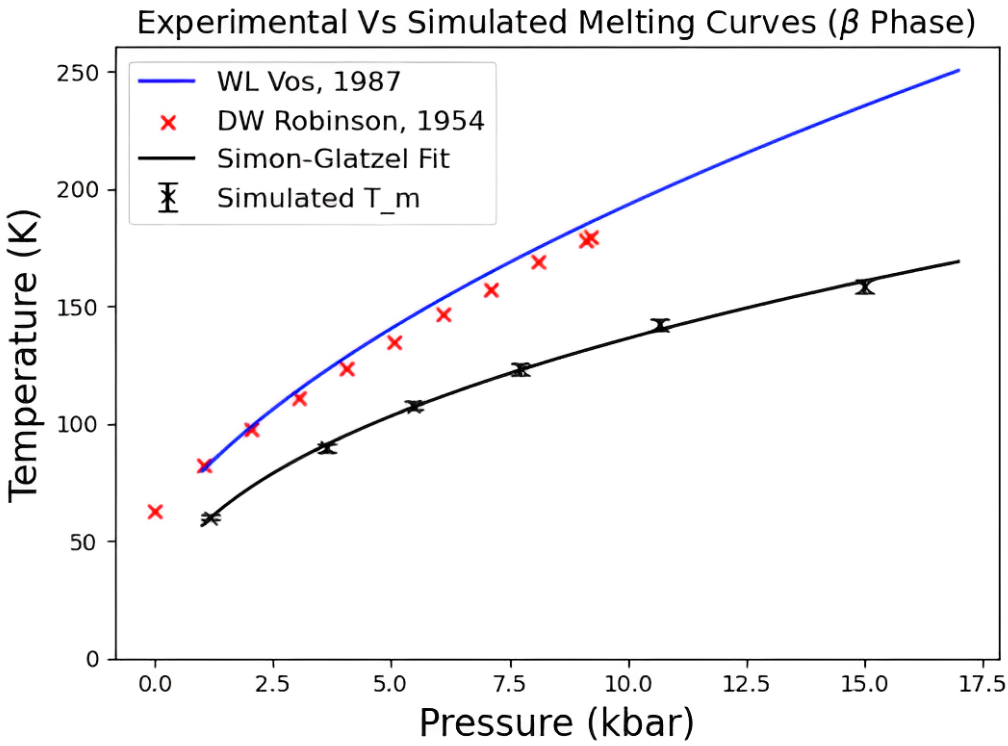


Figure 6.2: Computationally obtained nitrogen melt curve using Z-method and machine learned interatomic potential described in section 5.3. The model underestimate melting temperature by approximately 30 % as compared with the experimental data from [150, 199].

computation of the melt curve is the β N_2 phase. Its structure can be described by freely rotating molecules located at hcp lattice sites. The experimental lattice constant is 3.861 Å [155] therefore simulation box is scaled in the range from 3.84 to 4.15 Å. The simulation box contained around 1000 molecules which is generally more that sufficient for the method of choice. The Z-method simulations are performed using LAMMPS and interatomic potential presented in section 5.3. The molecules trajectories are simulated in the microcanonical ensemble using a time step of 1 fs. The total simulation time is 200 ps with the last 50 ps used to obtain time averaged temperature and pressure.

The obtained melt curve is shown in fig. 6.2 and compared with the experimental data from [150, 199]. The curve shows qualitatively agreement with experimental results, that is the increase in the melting temperature with pressure and the negative curvature of the line. However the model underestimates the melt be approximately 30 % and its predictions are clearly getting worse as pressure goes up. This might indicate that model will not be able to produce correct solid phases given pressure-temperature condition.

This discrepancy is likely to come from two independent sources. First, the Z-method might require extremely long simulation times when used for molecular systems (see [63] for more details). This is to ensure that the complete melting has occurred. Second, the potential model may not be accurate enough in this region of P, T space. Usually, a too-soft core region lowers the melting point, because it allows a wider range of molecular distances and thereby higher configurational entropy of the liquid [132].

Given that no solid data were used during the development of the model and its strong dependence on the density it is perhaps surprising that it produced even qualitative agreement with the melt. The model was designed to study fluid and its predictions in the solid region should be taken with a pinch of a salt. On the other hand it shows a limit on model's applicability and potential trap for unaware user.

In conclusion, the version of the potential from section 5.3 should not be used to study solid phases of nitrogen and its predictions are only valid in the liquid region where it was fitted. However, the qualitative agreement and quantitative disagreement suggests that the structure of the ML process is sound, and the model can be improved by fitting. This is addressed in the following section.

6.2 Improved N_2 Model Development

The first iteration of the model (section 5.3) successfully described subcritical and supercritical fluid regions on the P, T phase diagram. However it was deemed inadequate to study solid phases of nitrogen. The inability of the previous model to render the melt curve in proximity of the experimental results can be attributed to the following factors

- Many-body descriptor is density driven but no solid phases were used during its development.
- N_2 relative energy differences between competing phases can be lower than 1.0 meV/molecule requisite extremely accurate model.

In this section a new interatomic potential suitable for the molecular dynamics simulations is developed with greatly improved transferability as compared with the previous model. The model aims to accurately describe four-dimensional

potential energy surface (PES) of two interacting N_2 molecules with a simple rigid bond.

A new methodology is developed which is general in nature and can be applied to any dimer-like system as long as accurate training data are available. The method used in this section allows to build simple, high quality models directly from quantum chemistry calculations.

There are two major areas of overlap between the new model and the one developed in section 5.3. First, the same CCSDT(Q) data are used [85] during the training process, but this time the model is trained directly on the quantum chemistry data. The reader is referred to section 5.3.1 for details of the quantum chemistry calculations and how the training data set is constructed for the first model. Second, the same functional form is used to describe the local atomic environment of each atom as in the original model (section 5.3.1).

However, there is a significant difference in how atomic descriptor is utilised. Note, that the notation changes in this section (fig. 6.3). Here i_1 and i_2 are two bonded atoms of molecule i and similarly j_1 and j_2 belong to molecule j. Instead of allowing summation over all neighbouring atoms of the central atom i_1 within the cutoff distance r_c, the summations is restricted to just four distances: i_1-i_2, i_1-j_1, i_1-j_2 and j_1-j_2. Also the cutoff distance changes the meaning as compared with the first model. Instead of describing maximum distance between two interacting atoms, it is now a maximum distance between two molecule centres of mass. Therefore the local energy of molecule i is obtained by iterating over its all nearest neighbouring molecules and summing over each molecule-molecule interaction individually. The total energy of the system is then obtained by the summation of local molecular energies. Note that the forces are still computed between individual atoms. However, the force between bonded atoms is removed by the SHAKE algorithm [152] as implemented in LAMMPS such that the bond length is kept fixed at 1.1014 Å.

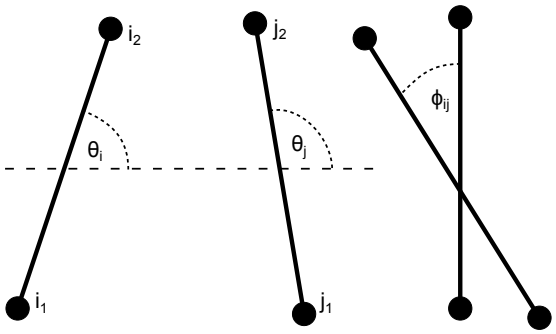


Figure 6.3: Schematic diagram representing internal coordinates of the N_2 molecule pair. Figure adapted from [85].

To allow molecular dynamics with this aforementioned methodology a new LAMMPS interface has had to be developed. Here, the major challenges were to prune atomic nearest neighbour lists such that each molecule-molecule interaction is counted just once and forces and stresses are accumulated accordingly. Our implementation differs to one already available in LAMMPS such that we only compute interaction energy if the molecular centres of mass are within the cutoff distance while LAMMPS operates on atom-atom distances.

Note that even though the model uses a rigid bond (enforced with the SHAKE algorithm) the i_1-i_2 and j_1-j_2 interactions are still calculated, so allowing the future model an additional two degrees of freedom for bond breaking. However, this approach would require extensive quantum chemistry data set being available for training. Alternatively, one can simply replace the rigid bond with a harmonic one. While in principle this approach might work, at the temperatures considered here the vibrational mode is in its quantum ground state. The bond is also very stiff, so the main effect of a flexible bond would be to introduce a spurious contribution to heat capacity in classical simulation.

Furthermore, the big difference in energy scales could cause numerical instability, so the justification of such a model can only be obtained by benchmarking it against relevant experimental data.

The PES of two interacting rigid dimers is four dimensional (6.3). One dimension describes the centre of mass distance between two interacting molecules i and j. Three dimensions describe their relative angular geometry: θ_i and θ_j are angles between the bond axes and the axis joining their centres of mass. The ϕ_{ij} is the dihedral angle between two molecular axes.

Fig 6.4 shows molecule-molecule interaction energies for various angular orientations as predicted by new MLIP. It is worth noting the differences in interaction energy between distinct angular arrangements at the same centre of mass distance. Notice that at the same distance the interaction between two molecules can be either attractive or strongly repulsive, depending on the molecular orientation. The obtained energy curves show excellent fit when compared with CCSDT(Q) data used for training. Majority of the predictions are within 0.1 meV/molecule. The largest deviation (0.35 meV/molecule) is observed for the lowest energy curve (black line) which corresponds to both molecules being tilted by 45° in the same direction from the axis joining their centres of mass. The strongest repulsion corresponds to collinear angular orientation where molecules

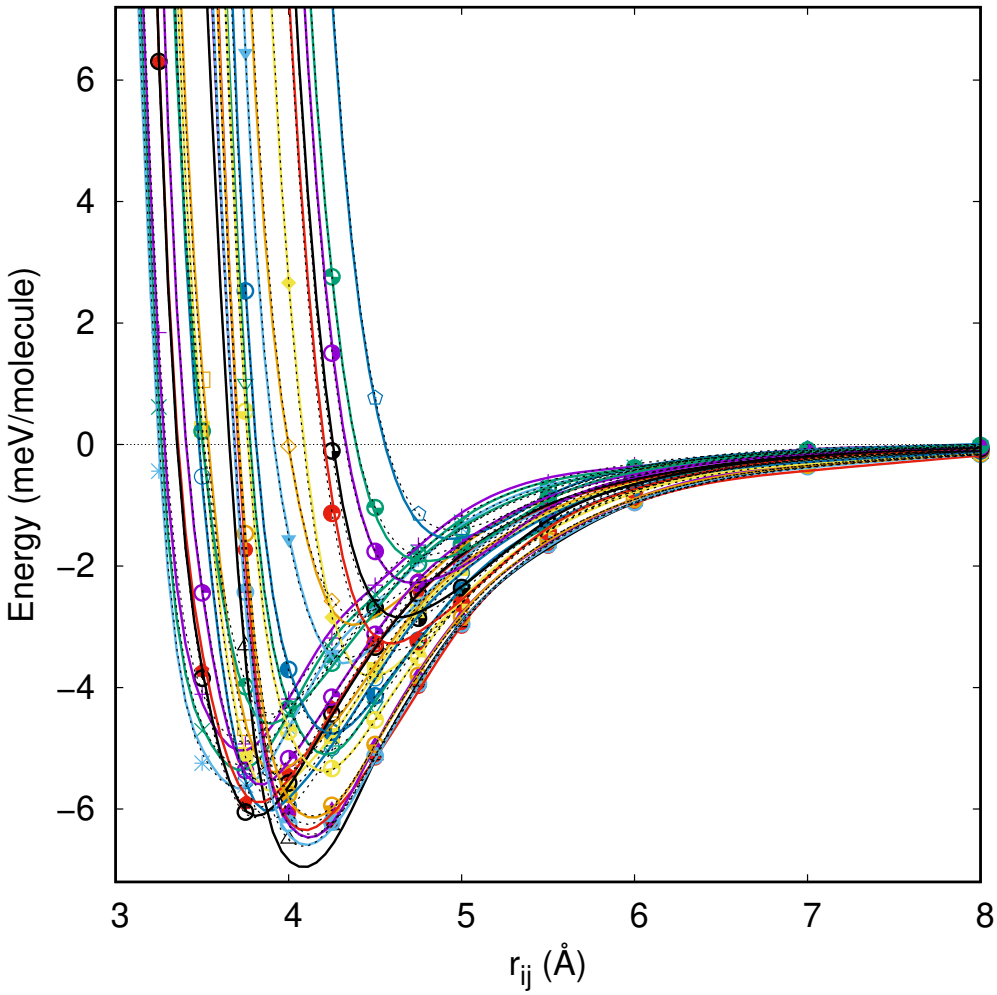


Figure 6.4: The interaction energy between the N_2 molecule pair as a function of molecule centre of mass distance for all 26 angular configurations (sets of θ_i θ_j, ϕ_{ij}) used during the training process. The CCSDT(Q) training data are represented by symbols. The fitted MLIP predictions are shown as lines. The dashed lines are from 5-centre model of reference [85]. Both models show excellent fit to CC data.

begin to repel each other just below 5 Å.

6.3 Nitrogen solid phases

Perhaps surprisingly, given simplicity of the nitrogen molecule, its experimental phase diagram is very rich (fig. 6.6). Just below 10 GPa six solid phases has been experimentally observed and their crystal structures identified. They are revived in this section as they are most relevant to this work. The low-temperature and low-pressure solid phases of nitrogen are of particular interest to current studies of planets in the outer Solar System. Fig. 6.5 shows image of the surface of the Pluto obtained by the New Horizon spacecraft. There is a striking difference between high water-ice mountains and nearly flat surface of nitrogen-rich ices. The strong hydrogen bond in water is sufficient to assemble icebergs while nitrogen-rich ices are rather soft an unable to support the weight of a mountains.

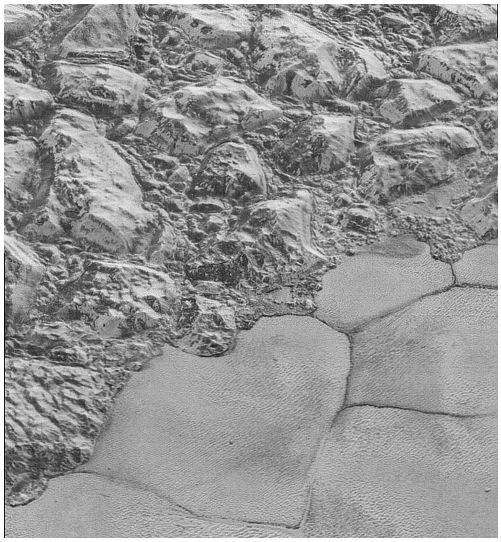


Figure 6.5: Image of the surface of Pluto obtained by the New Horizon spacecraft. Image shows water-ice mountains known as al-Idrisi. The mountain range ends abruptly at the Sputnik Planum where nitrogen rich ice forms almost level surface. The image is about 50 miles in width. Credit: NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute

.

First successful solidification of solid nitrogen was performed by Olszewski in 1884 [158]. The observed phase is now known as the β phase (see 6.3.2). In 1916 Euken [59] observed anomaly in the heat capacity upon further cooling of nitrogen at low pressure. The anomaly was correctly attributed to the phase transition from β to α phase (see 6.3.1) around 35 K at ambient pressure. In 1955 Swenson [179] obtained another phase (γ, see 6.3.3) by compressing α N_2 to 0.35 GPa at low temperature. Year later Stewart compressed β phase even further and observed the same γ phase but this time at 1 GPa and much higher temperate of 65 K. In 1984 Schiferl, Buchsbaum and Mills compressed γ phase at 15 K to obtain new ϵ phase (also known as δ(LT1)) [153] above 1.9 GPa (see 6.3.6). Two structures related to the ϵ phase can be obtained by either heating ϵ phase or by compressing β at room temperature (see δ 6.3.4 and δ^* 6.3.5) - first observed in 1979 as a split

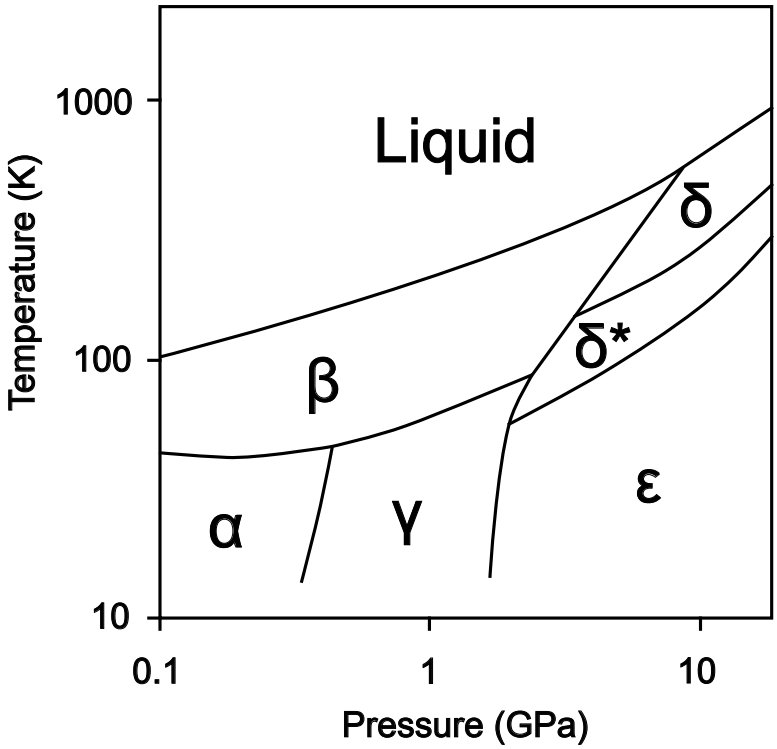


Figure 6.6: Experimental phase diagram of nitrogen. Figure adapted from [75]. The melting curve is from [208]. Greek letters label nitrogen solid phases. See text for phases description.

into a second high-frequency peak in the Raman line during compression of the β phase [111].

In the remaining part of this section these phases are described in more detail.

6.3.1 Alpha

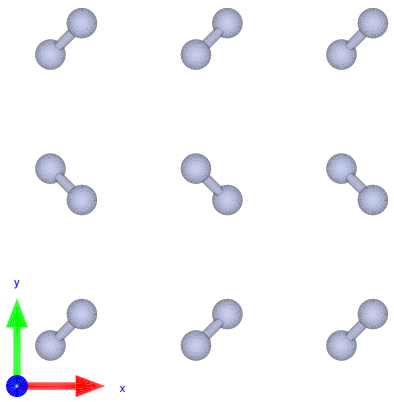


Figure 6.7: Low T and P ordered $Pa3$ cubic α phase.

Figure 6.7 shows alpha nitrogen (α-N_2) which is a low temperature and low pressure phase. Its crystal structure has been proposed by Ruhemann in 1932 as a $Pa3$ space group and finally confirmed from electron diffraction pattern in 1974 [197]. The unit cell contains four molecules each centred on a face centred cubic (fcc) lattice. Each molecule is aligned along a different cube body diagonal which preserves cubic symmetry. The experimental

lattice parameter is 5.433 Å at 0.3785 GPa and 19.6 K [155]. The ordering of molecules in the α phase is governed by the strong electric quadrupole-quadrupole interactions [49].

On heating the α phase transforms to β around 35 K, while when compressed it transitions to the γ phase above 0.35 GPa with slightly higher transition pressure at elevated temperatures.

6.3.2 Beta

The β phase is a dominant high temperature phase and is contiguous with a melting curve from zero up to approximately 10 GPa. It is a hexagonal $P6_3/mmc$ structure with two molecules per unit cell. The lattice parameters are $a = 3.861$ Å and $c = 6.265$ Å at 0.4125 GPa and 49 K [155]. The ratio of lattice parameters, c/a, is very close to the one obtained for hexagonal close-packed hard spheres - $\sqrt{8/3}$. The c/a ratio shows very little variation with temperature and pressure from the *ideal* one. This indicates that atomic positions are highly disordered with no evidence of ordering at higher pressures [155].

6.3.3 Gamma

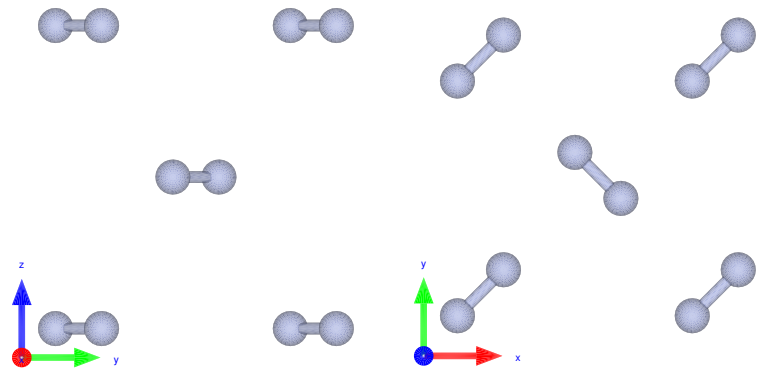


Figure 6.8: Body centred tetragonal γ phase with two molecules per unit cell.

The γ phase is a low temperature and moderate pressure ordered phase of nitrogen and is shown in fig. 6.8. Its crystal structure has been determined by x-ray diffraction as tetragonal with two molecules per unit cell at special position f of space group $P4_2/mnm$ [155]. Equivalently, it can be described as a body centred tetragonal (bct) lattice with a central molecule pointing along (110) direction and

the corner molecule being at orthogonal position $(\bar{1}10)$ to the central one. The unit cell parameters are $a = 3.957\,\text{Å}$ and $c = 5.109\,\text{Å}$ at an average pressure of 0.4015 GPa and average temperature of 20.5 K [155].

6.3.4 Delta

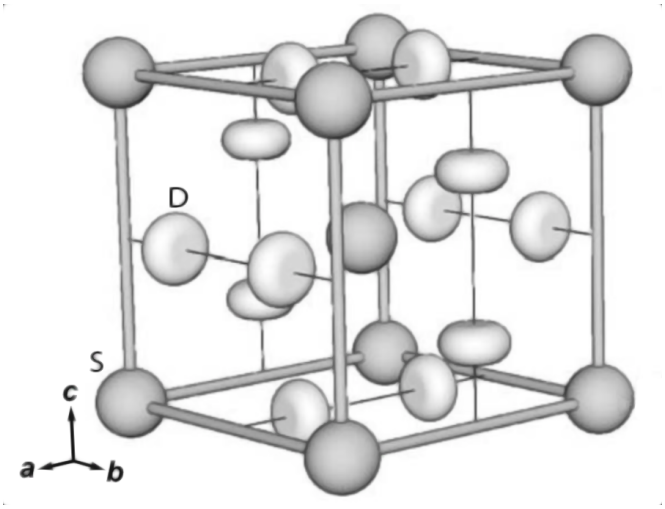


Figure 6.9: Unit cell of nitrogen δ phase. The cubic cell contains eight molecules. The central and corner molecule show nearly perfect spherical symmetry (labelled S on the diagram). The remaining six molecules have disk-like disorder (labelled with D).

The cubic δ phase has space group $Pm3n$ with eight molecules per unit cell and is similar to $\gamma - O_2$ and $\beta - F_2$ at 50 K and atmospheric pressure [45, 126]. The molecules show two distinct types of disorder. There are two molecules located at $2a$ Wyckoff sites at $(0, 0, 0)$ and $(1/2, 1/2, 1/2)$. Those molecules are almost perfectly spherically disordered however avoid pointing along the cubic $\langle 100 \rangle$ directions. The remaining six molecules are located at $6d$ Wyckoff sites at $(0, 1/4, 1/4)$ and the respective cubic symmetry equivalents. Their motion is disc-like with a uniform distribution of orientations. The lattice parameter at 5.7 GPa and 293 K is $a = 6.112(4)$ Å [175].

6.3.5 Delta*

Upon further compression of the cubic δ phase the closely related tetragonal δ^* is obtained. Its space group has been proposed as $P4_2/ncm$ in 1998 by [81]

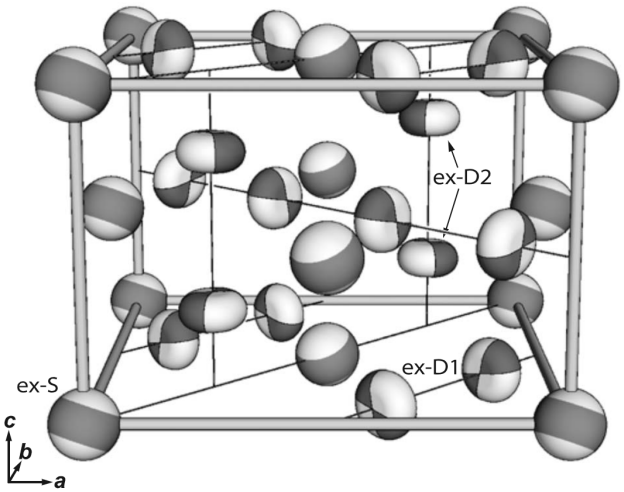


Figure 6.10: Schematic diagram of the tetragonal unit cell of δ^* phase which is closely related to the cubic δ phase. Here, almost perfect spherical disorder of the δ becomes slightly hindered and molecules begin to show preferred orientation (labelled *ex-S* where the dark area indicates favoured molecular direction). Similarly, the disk-like motion is no longer uniform (*ex-D1* and *ex-D2*). Reprinted from [175], with the permission of AIP Publishing.

and finally resolved in 2009 by [175]. The measured unit cell parameters are $a = 8.063(5)$ Å and $c = 5.685$ Å at 14.5 GPa and 293 K.

The diffraction pattern of the δ and δ^* phases are very similar. Moreover, the δ^* phase is an intermediate phase between fully ordered ϵ phase and almost perfectly disordered δ phase. It was found that the molecular centres positions of the δ phase are the same as in the rhombohedral ϵ phase [126]. It is therefore perhaps unsurprising that the δ^* shares the same positions for the molecular centres [175]. However, in the δ^* phase the molecular orientation disorder is reduced and molecules appear to show preferred directions. The refinement of the structure has been performed in [175] and the resulting unit cell is shown in 6.10. Unfortunately, the proposed structure does not produce the observed number of Raman and infrared modes as pointed out by the authors. As compared with the δ phase the sphere-like molecules prefer aligning along $\langle 3\bar{1}\sqrt{2}\rangle$ directions and avoid $\langle 01\sqrt{2}\rangle$ direction but still shows spherical disorder. There are two sets of disk-like molecules which show coordinated motion where molecular orientations are either paired or perpendicular to each other [175].

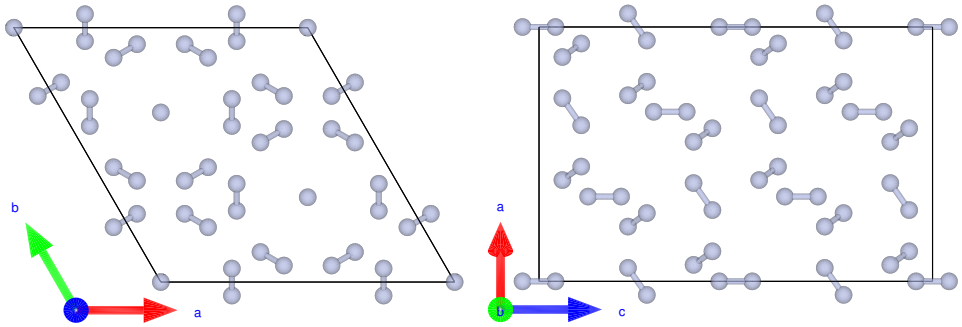


Figure 6.11: Rhombohedral unit cell of nitrogen ϵ phase.

6.3.6 Epsilon

The rhombohedral ϵ phase is best described as a small distortion to a cubic δ phase along $\langle 111 \rangle$ direction - the resulting angle between axes is around 5° [126]. The molecular centre positions are slightly displaced as compared with the δ and δ^* phases and without spherical- or disc-like disorder. Similarly to the α and γ low temperature phases, molecular motion is restricted to libration. The space group has been established as $R\bar{3}c$ and it remains stable in approximately 2 to 25 GPa range [126, 153].The rhombohedral unit cell contains eight ordered molecules. The similarity between δ, δ^* and ϵ is apparent from their respective Raman stretching-mode spectra [153] each containing distinct two branches - intense low frequency peak and less pronounced higher frequency peak. The ϵ phase can either be obtained by compressing γ phase at low temperature or δ phase at room temperature [135]. The hexagonal unit cell dimensions at 16.3 GPa and room temperature are $a = 7.605$ Å and $c = 10.622$ Å. The c/a ratio tends to increase with pressure from 1.396 at 16.3 GPa to 1.427 at 43.9 GPa [135].

6.4 Nitrogen Phase Diagram

The improved model for molecular nitrogen from section 6.2 is employed to compute the phase diagram of nitrogen up to 10 GPa.

There are six relevant phases as reviewed in section 6.3. The molecules in low temperature phases (α, γ and ϵ) are ordered, and their motion is libration (molecule oscillates back and forth). The high temperature phases (β, δ and δ^*) are either partially or fully disordered rotors (3D rotation around the fixed centre of mass) with molecules showing complex sphere- or disk-like motion. While our

focus is on the experimentally observed phases it is worth noting that the methods employed in this chapter might be used to identification of new crystal structures for N_2. In particular NPT simulations allow for spontaneous phase transition to occur. However, given relatively short simulation times, as compared with the experiment, it is usually the case that the transition can happen only in the absence of a high energy barrier. Moreover, the fixed number of molecules in the simulation cell will favour phase transition to phases with the commensurate number of molecules in the unit cell. So, new crystal structures appearing on our simulations are treated with caution when they are similar to the preceding phase.

6.4.1 The melting curve

The phase coexistence calculations are performed to obtain the melt line of molecular nitrogen. The phase coexistence is a well established and very accurate method albeit computationally more expensive than so-called Z-method [20, 21].

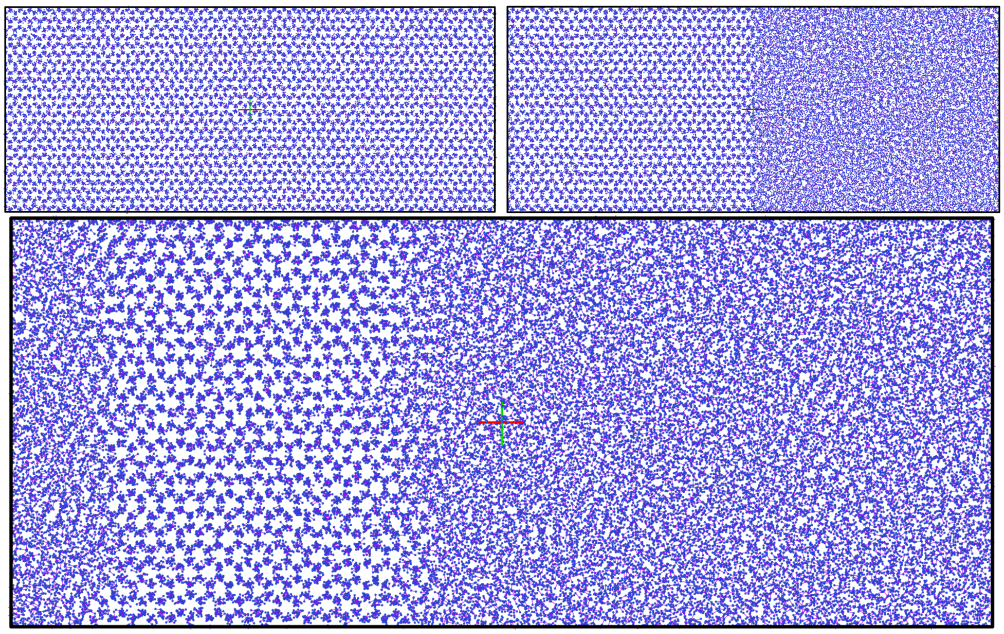


Figure 6.12: Snapshots taken from the molecular dynamics run showing progressive stages of the phase coexistence method. The top left figure shows equilibrated nitrogen β phase. In the top rigth figure atoms in the left hand side of the box are frozen while the remaining atoms are heated up above the melt point. The bottom image shows equilibrated phase coexistence of the solid β phase and liquid.

The procedure used here for phase coexistence calculations can be summarised

as follows. The initial box contains the relevant solid phase for a given pressure. The initial configuration is equilibrated for 20 ps in the NPT ensemble with temperature and pressure being close to the expected melt point. The timestep of 1 fs is used throughout. After initial equilibration, approximately half of the box is kept frozen while the remaining molecules are first heated to $1.5T_m$, where T_m is experimental melting temperature, then cooled down to the initial temperature. Melting and cooling takes 10 ps each. Finally the NPH ensemble is used to simulate entire system for at least 350 ps. There are three possible scenarios at this stage. The molecules in the box either complete solidify, melt or a mixture of solid and liquid is present at the end of the simulation. The first two cases indicate that the initial temperature was too low or to high respectively. The latter case is desired because it means that the simulation has equilibrated at thermodynamic pressure and temperature conditions somewhere on the melt curve. The time averaged instantaneous kinetic energy from the last 50 ps is assumed to be the melting temperature. By changing the pressure, it is possible to track the entire melt curve.

The β and δ phases have been experimentally determined as being adjacent to the melt curve up to 10 GPa. The simulation box of the β phase contains 28800 molecules on the hcp lattice. The molecular orientations are assigned at random and show full spherical disorder/rotation throughout the simulation. The initial structure for δ has space group $Pm3n$. The cell is constructed with 29952 molecules. The molecular orientations where randomly assigned. It is observed that at temperatures below the expected melting temperature nitrogen molecules located at $6d$ Wyckoff sites show disk-like motion while the remaining ones are spherically disordered. The Wyckoff positions represent equivalent sites in a crystal which are linked by the group symmetry operations.

P (GPa)	β T (K)	δ T (K)
0.1	83	-
0.3	112	-
0.5	142	-
1.0	185	-
2.45	297	-
5.0	389	365
7.0	440	433
8.0	457	458
9.0	467	474
10.0	488	496

Table 6.1: The melting temperatures obtained from the phase coexistence simulations between 0.1 GPa and 10.0 GPa for both β and δ phases of nitrogen.

The total of 15 coexistence calculations (10 simulations begin with the β phase and 5 as the δ phase) were performed to obtain smooth melt line shown in fig.

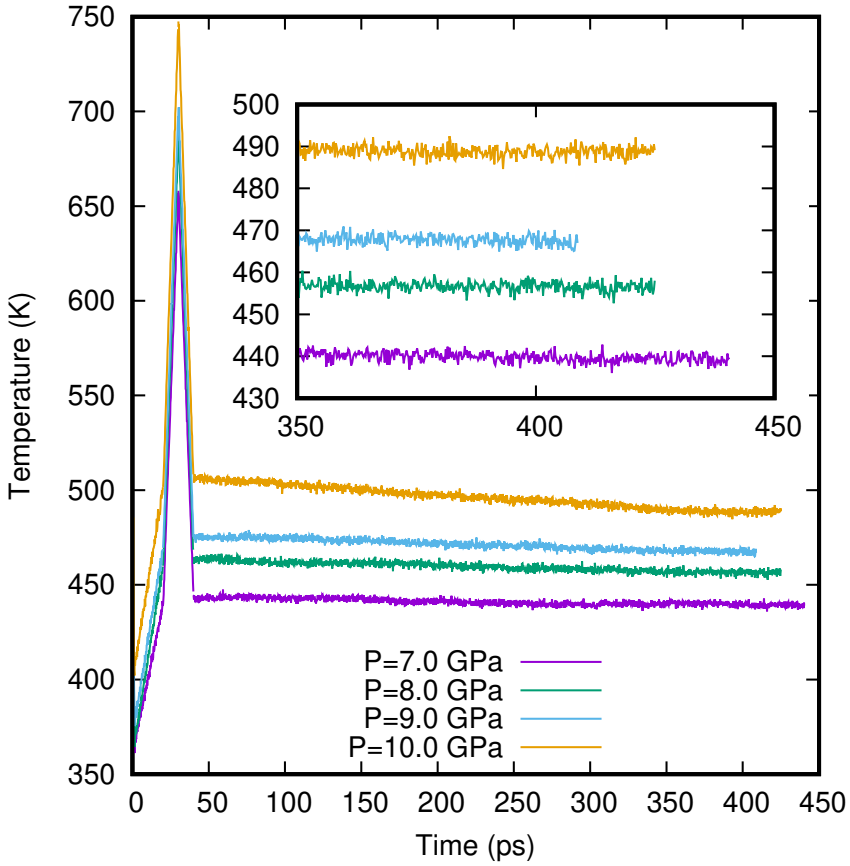


Figure 6.13: The figure shows temperature vs time for phase coexistence MD simulations of the β phase at various pressures. The sharp peak in temperature around 30 ps corresponds to melting of the half of the simulation box. The last 50 ps were used to obtain the melting temperature (see inset).

6.14 as a solid red line. The temperature against the time plots are shown in fig. 6.13 for a number of different pressure points. The initial guess for a melting temperature is critical for a successful application of the phase coexistence method. The calculated melting temperatures under various pressures are shown in table 6.1. The obtained melt curve is in a very good agreement with the experimental one as shown in fig. 6.6.

Here, we also compare the performance of the MLIP model to the 5CM model. The 5CM model is approximately an order of magnitude faster with respect to MLIP even though it make use of a 16 % longer cutoff distance. The 5CM model utilises the optimised tabulated implementation of pair potentials while MLIP does not. The 5CM model requires computation of 25 pairwise functions for each molecule-molecule interaction while MLIP uses just 4 distances to compute 4 pairwise functions and 4 densities. Therefore, we speculate that the tabulated implementation of MLIP might be on par with the 5CM model. We do not

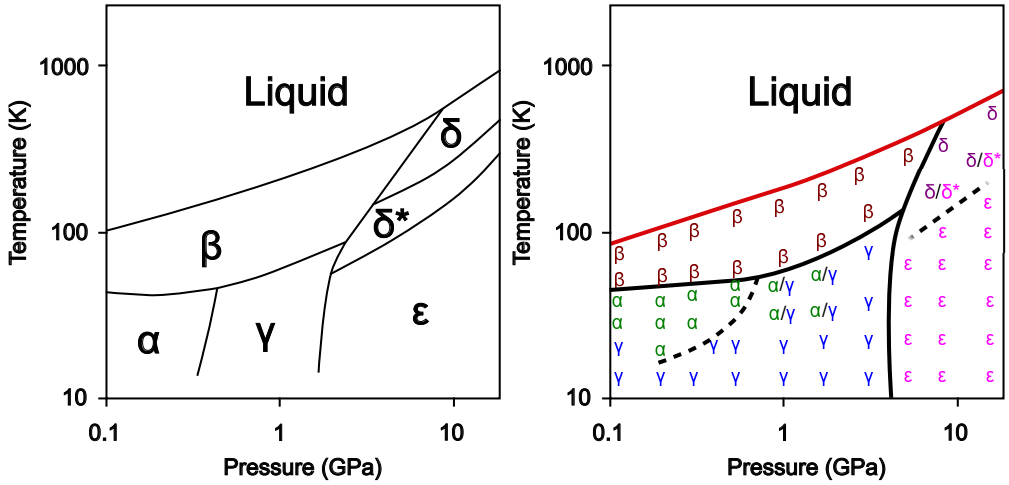


Figure 6.14: (Left) Experimental phase diagram. The same as in fig. 6.6; repeated here for convenience. (Right) The phase diagram for molecular nitrogen up to 10 GPa obtained using improved MLIP from section 6.2. Greek letters are used to label N_2 solid phases as explained in the text. The melt curve (solid red line) is computed from phase coexistence simulations. Solid lines are obtained phase boundaries while dashed lines indicate likely transitions. See text for more detail. The computed phase diagram is in very good agreement with the experimental phase diagram (fig. 6.6).

run any tests with standard Lennard-Jones potential. The two-site LJ potential requires computation of four pairwise functions therefore we can estimate it to be approximately five-times faster (25/4) than 5CM model.

6.4.2 Solid phase boundaries

The phase boundary between $\alpha - \beta$ and $\gamma - \beta$ is represented on the phase diagram in fig. 6.14 as a solid black line with an increasing positive slope with pressure. The phase boundary has been estimated based on the following observations. Near the experimental boundary it is observed that during the molecular dynamics simulations across the boundary, molecules in the α phase start to rotate or equivalently β rotors cease their motion. This behaviour is only observed in the narrow temperature zone, and we take it as indicating the position of the phase boundary between librating and rotating molecules.

However, we note that no full phase transition is obtained from fcc to hcp or vice versa. The molecular centres in simulations started in the β phase remain hexagonal even after rotation ceases; similarly the molecular centres of α remain close to fcc. In both case, the rotation ceases at similar T,P conditions.

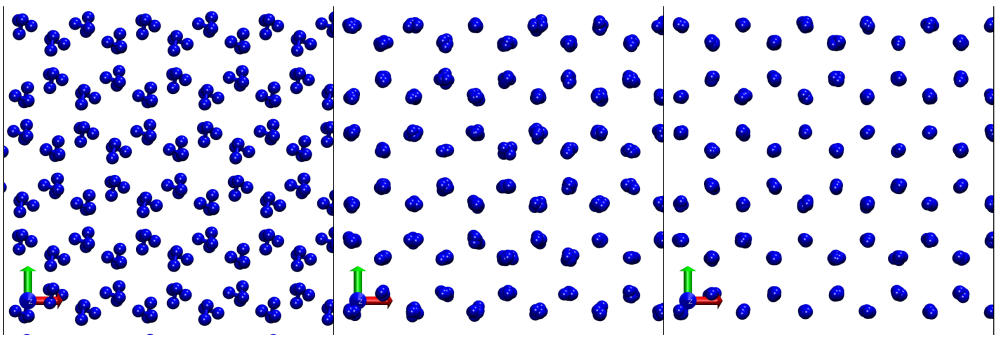


Figure 6.15: Molecular dynamics snapshots of the β phase time averaged over 200 steps at three different temperatures and P=0.3 GPa. From left to right: T=30 K, T=40 K, T=50 K. The molecular motion progressively changes from librons to rotors on heating. The estimate phase boundary is just below 40 K in agreement with the experiment. Note that molecular centres remain always on β phase hcp sites. The time averaged rotor position results in a single point on the lattice site.

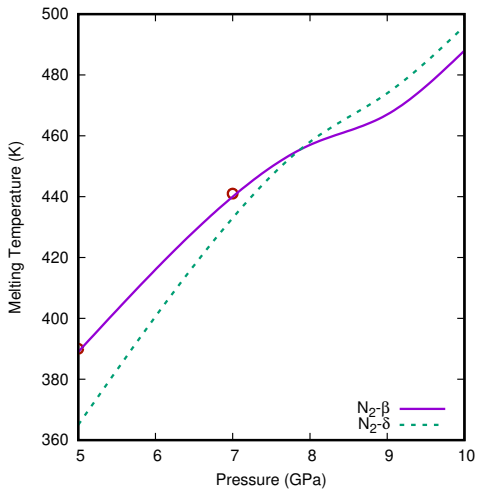


Figure 6.16: The melting temperature of the β and δ phases. The crossover of the melting curves at 7.9 GPa indicates position of the β/δ/liquid triple point. Red circles are computed with 5CM model from [85].

This is understandable as even for an atomic system fcc/hcp phase transition is complex and difficult to realise in molecular dynamics simulation. For example, the hcp to fcc transition in titanium is a process which involves slip of planes dislocations, adjustment of interplanar spacing followed by the volume expansion [205]. To best of my knowledge the mechanism behind the fcc/hcp phase transition for N_2 (or any other) dimers is unknown.

If we consider the static hcp and rotating fcc phases to be metastable, then the heating simulations give and upper bound on the true phase line, while the cooling calculations give a lower bound: this enables the phase boundary to be determined with high confidence.

Upon heating of the tetragonal γ phase, there is a small increase in the c/a ratio. Close to the phase boundary the molecules begin to rotate and there is a sudden jump in the c/a ratio to $\sqrt{2}$ indicating a transition to a perfect fcc lattice. The

initial transition from body centred tetragonal lattice to fcc follows classical Bain transformation [9]. The complete phase transition to the hexagonal β phase does not occur due to the same reason as for $\alpha - \beta$ as explained above.

The triple point between β/δ/liquid is obtained directly from the crossover of the melt curves computed with respective structures as shown in fig. 6.16. The triple point is located at 8 GPa which is in very good agreement with the experimental one which is around 9 GPa [192]. The positive slope for the $\beta - \gamma$ boundary was initially motivated a priori based on the experimental evidence. However it is reasonable to infer that the high temperature triple point should smoothly join with the zero temperature $\gamma - \epsilon$ boundary as explained below.

A zero temperature point on the boundary between γ and ϵ phases is obtained from the relative enthalpy differences. Fig. 6.17 (left) shows computed enthalpies. The γ phase becomes dominant above 4 GPa at zero temperature. The experimentally determined phase transition is around 2 GPa at 15 K [153] indicating that the model either overestimates the transition pressure or the phase boundary between γ and ϵ has changing slope from negative slope at low pressure to positive at temperatures above 20 K. The zero temperature boundary point joins smoothly with the β, δ, liquid triple point around 8 GPa. This behaviour is expected due to the similar crystal structures of the δ and ϵ phases.

On heating from ϵ above 8 GPa to is observed that the molecular motion changes and molecular orientations progressively evolves libron-like motion to spherical rotor or disk-like. The simplified data analysis used in this section is inadequate to establish two distinct boundaries between $\epsilon - \delta^*$ and $\delta^* - \delta$ phases. It is known that both δ structures are closely related and the main difference is the molecular orientations (see 6.3.4 and 6.3.5).

The dashed line separating δ and and ϵ phases indicates the approximate position of the rotation transition.

It is observed that during MD simulations the α phase spontaneously transforms to γ at temperatures below 20 K across the computed pressure range. The phase transition occurs within the first few picoseconds of the simulation. This observation is further confirmed by comparing enthalpy differences at pressures below 0.5 GPa as shown in fig. 6.17 (right). The γ phase is marginally more stable at low pressure but its relative stability increases suddenly near the expected phase boundary at 0.3 GPa. The obtained free energy differences at zero temperature are around 1 meV/molecule in the measured pressure range.

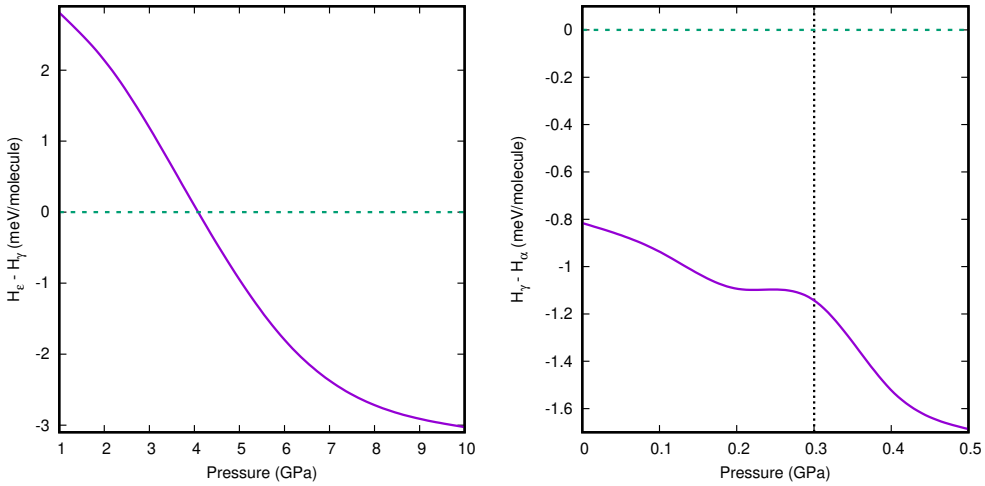


Figure 6.17: (Left) Enthalpy difference between the nitrogen γ and ϵ phases as a function of pressure. The ϵ phase becomes stable above 4 GPa and its relative stability further improves with pressure. (Right) Enthalpy difference between the nitrogen γ and α phases as a function of pressure. The γ phase is a dominant low temperature phase. The vertical line indicates approximate position of the experimental phase boundary.

The small enthalpy differences and observed phase transition indicates low phase transition energy barrier below 20 K. However, above those temperatures the phase transition is not observed indicating at least strong metastability of the α phase. The experimental evidence points to α being stable at low temperature and low pressure. The behaviour of the model is therefore different with the γ phase being the most stable low temperature phase.

The γ phase provides optimised packing for oblate objects, while α is fcc which provides optimal packing for spheres. As the temperature increases the libron motion rises and the oblate-like 2D rotor progressively becomes a sphere-like 3D-rotor for which fcc provides best packing ratio hence minimising free energy. The fcc phase allows for greater motion at the same temperature as compared with the tetragonal γ hence allowing it to generate higher entropy. It is therefore reasonable to expect, positive slope of the phase boundary between those phases.

Chapter 7

Conclusions and Future Work

Molecular dynamics (MD) simulations are a practical and convenient computational tool in the physical sciences. The applicability of this method strongly depends on the quality and computational efficiency of the interatomic potential. Machine learning interatomic potentials (MLIP) can provide an accurate description of the potential energy surface comparable with first principles methods such as density functional theory, or even coupled cluster methodology, at the fraction of computational cost.

This thesis highlighted development of *Ta-dah!* - a software package dedicated to construction of machine learning interatomic potentials. The software is capable of generating production-ready potentials which then can be directly deployed to run molecular dynamics simulations. The design of the package allows implementation of new methods, such as descriptors or optimisers, with a minimum effort. The universal LAMMPS plugin works with new methods seamlessly, thereby dramatically reducing the time between development, testing and deployment of the model. The *Ta-dah!* package attempts to tackle the problem of transferability of MLIPs by introducing two-stage fitting procedure. The software in interfaced to standard DFT codes, and also provides a novel method which allows effortless development of MLIPs for dimer-like systems using existing quantum chemistry data.

The *Ta-dah!* is employed to develop accurate MLIP for molecular nitrogen using publicly available couple cluster data. The model is benchmarked against experimental data and deployed to study the Frenkel Line (FL) - the transition between gas-like region and liquid-like region in the supercritical fluid. The model

predicts termination of the FL at the triple point. This model is further refined to study the phase diagram of nitrogen up to 10 GPa. Given the simplicity of the model it is remarkable that it is capable of reproducing all solid phases of nitrogen including accurate position of the melting line.

This thesis presented particular case for molecular nitrogen however, the *Ta-dah!* package is capable of developing MLIPs for many different classes of materials such as metals or insulators. The *Ta-dah!* has been successfully applied within the School to develop potentials such as general purpose potentials for tantalum and krypton as well as preliminary model for titanium-niobium alloys. The aim of *Ta-dah!* is to promote development of new interatomic potentials as well as to accelerate development of new methods for descriptors, regression models or a combination of both.

The code is written in C++ and provide an easy-to-use command line interface. However the usage of the code as a library requires some basic knowledge of C++. This language was chosen for its computations efficiency and flexibility in expressing programmers ideas. Arguably, C++ is not a top choice for scientists. It should be relatively easy to develop python API to cover relevant functionality hence increase its appeal to a wider community.

The main effort was directed into developing a flexible framework and not providing the widest possible range of descriptors. Therefore, somewhat obviously, addition of new descriptors and optimisers such as neural networks should be prioritised in the future. Also, efforts should be directed towards the development of new descriptors for molecular systems such as water or methane.

As it stands, *Ta-dah!* does not support automatic ways to construct training data sets. Random data generation does not sample efficiently the configurational space and results in many redundant structures. This itself affects the efficacy of machine learning algorithms. The hand-building of data sets can be effective but extremely time consuming. It is only natural to direct future developments towards better automatic ways to sample the configurational space, focused on regions which are sampled by simulation with the potential, but not in the training set.

The *Ta-dah!* Nitrogen potential reproduces the five experimentally-known phases $\alpha-\epsilon$. Experiment has yet to identify the orientations for molecules in the δ^* phase. Given an accurate description of the phase boundaries and the melting curve the new potential can be employed to study this phase in detail. Furthermore, the

nitrogen potential can be employed to study the phase diagram above 10 GPa where new experimental phases are still being discovered.

The code for *Ta-dah!* is available at

`https://git.ecdf.ed.ac.uk/s1351949/ta-dah`.

and documentation at

`https://ta-dah.readthedocs.io/en/latest/`.

The documentation provides a quick start guide on how to use command line interface. The *Ta-dah!* git repository comes with a directory containing examples on API usage.

Appendix A

RDFs for N_2 MLIP - Model 1

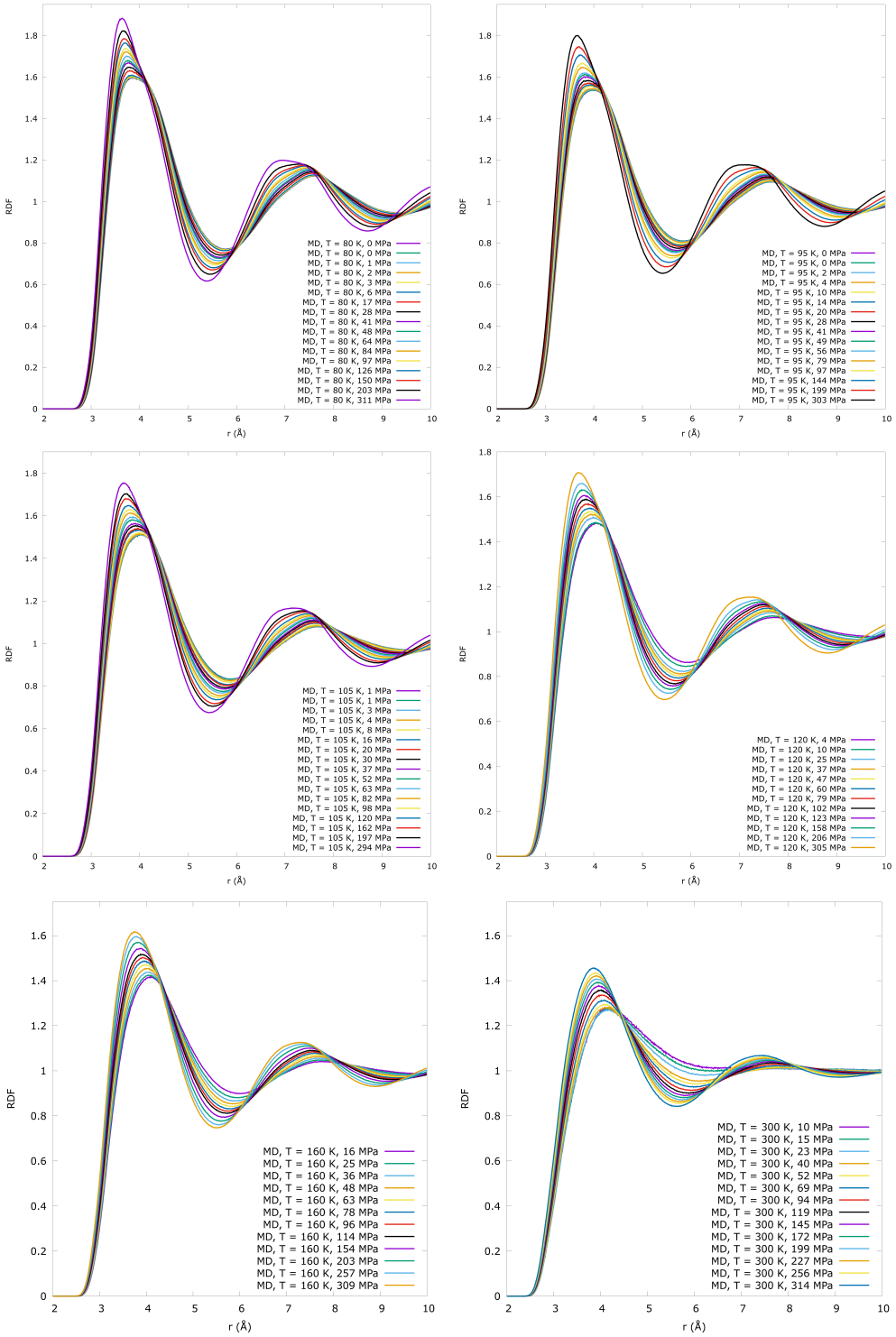


Figure A.1: Radial Distribution Functions for N_2 ML model for all subcritical and supercritical isotherms followed in the present study.

Appendix B

Coordination number for different sets of HPs - Model 1

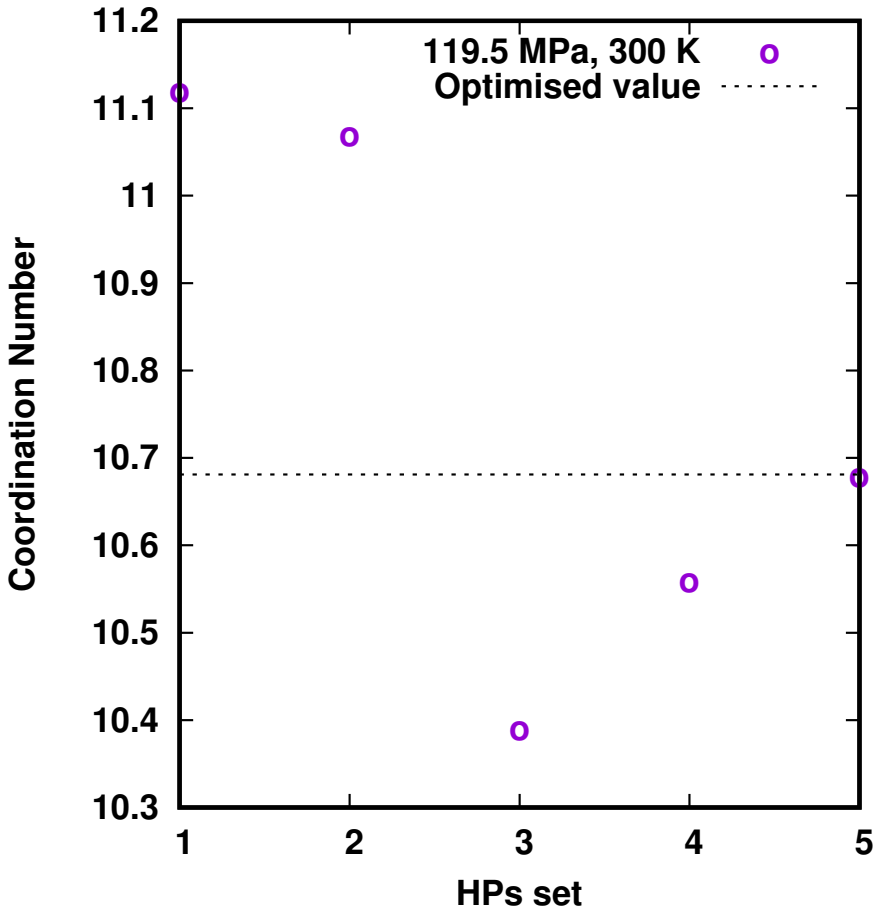


Figure B.1: The figure shows variation of the coordination number for the N_2 model developed in chapter 5. Each model is trained using the same training data set but a different set of hyperparameters (HPs). The HPs are optimised manually such that the pair correlation function resemblances the one obtained from the neutron diffraction experiments. The variation in CN during the optimisation process is within the experimental error in CN (see fig. 5.4).

Appendix C

Convergence of Model 2 with respect to the amount of training data

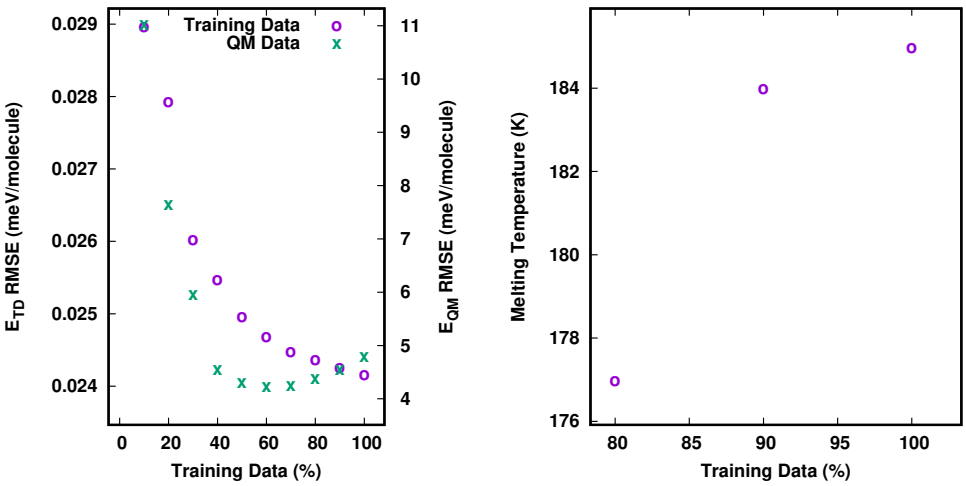


Figure C.1: The figure shows convergence of the model developed in chapter 6 with respect to the decreasing amount of training data. (Left) Energy root mean square error (RMSE) is computed for training data (TD, shown as circles) and coupled cluster data (QM, shown as crosses). The QM data set contains very high energy configurations (over 1 eV) resulting in a relatively large RMSE. The increase in RMSE for QM data above 60 % is a result of model's emphasis on accurate PES description around the equilibrium distance. (Right) Melting temperature at 1 GPa for models trained with decreasing amount of data. It was found that models trained with less than 80 % of data do not provide stable trajectories. We deduce that the additional data are required to prevent overfitting of the model and in consequence smooth potential energy surface.

Bibliography

[1] Ackland. "Semiempirical model of covalent bonding in silicon." *Physical review. B, Condensed matter* 40, 15: (1989) 10,351–10,355. http://www.ncbi.nlm.nih.gov/pubmed/9991580.

[2] Ackland, G. J., D. J. Bacon, A. F. Calder, and T. Harry. "Computer simulation of point defect properties in dilute Fe—Cu alloy using a many-body interatomic potential." *Philosophical Magazine A* 75, 3: (1997) 713–732.

[3] Ackland, G. J., G. Tichy, V. Vitek, and M. W. Finnis. "Simple ¡i¿N¡/i¿ -body potentials for the noble metals and nickel." *Philosophical Magazine A* 56, 6: (1987) 735–756.

[4] Ackland, G., A. Sutton, and V. Vitek. "Twenty five years of Finnis–Sinclair potentials." *https://doi.org/10.1080/14786430903271005* 89, 34-36: (2009) 3111–3116. https://www.tandfonline.com/doi/abs/10.1080/14786430903271005.

[5] Admal, N. C., and E. B. Tadmor. "A Unified Interpretation of Stress in Molecular Systems." *Journal of Elasticity 2010 100:1* 100, 1: (2010) 63–143. https://link.springer.com/article/10.1007/s10659-010-9249-6.

[6] Agrawal, A., and A. Choudhary. "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science." *APL Materials* 4, 5: (2016) 053,208.

[7] Artrith, N., A. Urban, and G. Ceder. "Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species." *Physical Review B* 96, 1.

[8] Axilrod, B. M., and E. Teller. "Interaction of the van der Waals Type Between Three Atoms." *The Journal of Chemical Physics* 11, 6: (1943) 299–300. https://pubs.aip.org/aip/jcp/article/11/6/299/182086/Interaction-of-the-van-der-Waals-Type-Between.

[9] Bain, E. C. "The Nature of Martensite." *AIME, Steel* 504: (1924) 70.

[10] Bartók, A. P., R. Kondor, and G. Csányi. "On representing chemical environments." *Physical Review B* 87, 18: (2013) 184,115. https://link.aps.org/doi/10.1103/PhysRevB.87.184115.

[11] Bartók, A. P., M. C. Payne, R. Kondor, and G. Csányi. "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons." *Physical Review Letters* 104, 13: (2010) 136,403. https://link.aps.org/doi/10.1103/PhysRevLett.104.136403.

[12] Baskes, M. I. "Modified embedded-atom potentials for cubic materials and impurities." *Physical Review B* 46, 5: (1992) 2727. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.46.2727.

[13] Baskes, M. I., J. S. Nelson, and A. F. Wright. "Semiempirical modified embedded-atom potentials for silicon and germanium." *Physical Review B* 40, 9: (1989) 6085. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.40.6085.

[14] Becke, A. D. "Density-functional exchange-energy approximation with correct asymptotic behavior." *Physical Review A* 38, 6: (1988) 3098. https://journals.aps.org/pra/abstract/10.1103/PhysRevA.38.3098.

[15] Becker, C. A., F. Tavazza, Z. T. Trautt, and R. A. Buarque de Macedo. "Considerations for choosing and using force fields and interatomic potentials in materials science and engineering." *Current Opinion in Solid State and Materials Science* 17, 6: (2013) 277–283.

[16] Behler, J. "Representing potential energy surfaces by high-dimensional neural network potentials." *Journal of Physics: Condensed Matter* 26, 18: (2014) 183,001. http://stacks.iop.org/0953-8984/26/i=18/a=183001?key=crossref.e2110f4e5f0e5ce5600d2eb9e27e4391.

[17] ———. "Perspective: Machine learning potentials for atomistic simulations." *Journal of Chemical Physics* 145, 17: (2016) 170,901. http://aip.scitation.org/doi/10.1063/1.4966192.

[18] Behler, J., S. Lorenz, and K. Reuter. "Representing molecule-surface interactions with symmetry-adapted neural networks." *The Journal of Chemical Physics* 127, 1: (2007) 014,705. https://aip.scitation.org/doi/abs/10.1063/1.2746232.

[19] Behler, J., and M. Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces." *Physical Review Letters* 98, 14: (2007) 146,401. https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

[20] Belonoshko, A. B., L. Burakovsky, S. P. Chen, B. Johansson, A. S. Mikhaylushkin, D. L. Preston, S. I. Simak, and D. C. Swift. "Molybdenum

at High Pressure and Temperature: Melting from Another Solid Phase." *Physical Review Letters* 100, 13: (2008) 135,701.

[21] Belonoshko, A. B., N. V. Skorodumova, A. Rosengren, and B. Johansson. "Melting and critical superheating." *Physical Review B* 73, 1: (2006) 012,201.

[22] Bender, C. M., P. N. Meisinger, and Q. Wang. "All Hermitian Hamiltonians have parity." *Journal of Physics A: Mathematical and General* 36, 4: (2003) 1029. https://iopscience.iop.org/article/10.1088/0305-4470/36/4/312https://iopscience.iop.org/article/10.1088/0305-4470/36/4/312/meta.

[23] Bengio, Y. "Gradient-based optimization of hyperparameters." *Neural Computation* 12, 8: (2000) 1889–1900.

[24] Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl. "Algorithms for Hyper-Parameter Optimization." *Advances in Neural Information Processing Systems* 24.

[25] Bergstra, J., and Y. Bengio. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research* 13, 10: (2012) 281–305. http://jmlr.org/papers/v13/bergstra12a.html.

[26] Bishop, C. M. *Machine Learning and Pattern Recogniton.* Springer-Verlag, 2006.

[27] Blank, T. B., S. D. Brown, A. W. Calhoun, and D. J. Doren. "Neural network models of potential energy surfaces." *The Journal of Chemical Physics* 103, 10: (1998) 4129. https://aip.scitation.org/doi/abs/10.1063/1.469597.

[28] Bloch, F. "Über die Quantenmechanik der Elektronen in Kristallgittern." *Zeitschrift für Physik 1929 52:7* 52, 7: (1929) 555–600. https://link.springer.com/article/10.1007/BF01339455.

[29] Blum, C., and A. Roli. "Metaheuristics in combinatorial optimization." *ACM Computing Surveys (CSUR)* 35, 3: (2003) 268–308. https://dl.acm.org/doi/10.1145/937503.937505.

[30] Bolmatov, D., M. Zhernenkov, D. Zav'yalov, S. N. Tkachev, A. Cunsolo, and Y. Q. Cai. "The Frenkel Line: a direct experimental evidence for the new thermodynamic boundary." *Scientific Reports 2015 5:1* 5, 1: (2015) 1–10. https://www.nature.com/articles/srep15850.

[31] Born, M., and R. Oppenheimer. "Zur Quantentheorie der Molekeln." *Annalen der Physik* 389, 20: (1927) 457–484. https://onlinelibrary.wiley.com/doi/full/10.1002/andp.19273892002https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19273892002https://onlinelibrary.wiley.com/doi/10.1002/andp.19273892002.

[32] Boys, S. F. "Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 200, 1063: (1950) 542–554. https://royalsocietypublishing.org/doi/10.1098/rspa.1950.0036.

[33] Brazhkin, V. V., Y. D. Fomin, A. G. Lyapin, V. N. Ryzhov, and K. Trachenko. "Two liquid states of matter: A dynamic line on a phase diagram." *Physical Review E* 85, 3: (2012) 031,203. https://link.aps.org/doi/10.1103/PhysRevE.85.031203.

[34] Buckingham, R. A., Buckingham, and R. A. "The classical equation of state of gaseous helium, neon and argon." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 168, 933: (1938) 264–283. https://royalsocietypublishing.org/doi/10.1098/rspa.1938.0173.

[35] Carlsson, A. E. "Beyond Pair Potentials in Elemental Transition Metals and Semiconductors." *Solid State Physics - Advances in Research and Applications* 43, C: (1990) 1–91.

[36] Carrasquilla, J., and R. G. Melko. "Machine learning phases of matter." *Nature Physics* 13, 5: (2017) 431–434.

[37] Cauchy, A.-L. "Mémoire Sur Les Deux Espéces D'ondes Planes Qui Peuvext Se Propager Dans Un Systéme Isotrope De Points Matériels." In *Oeuvres complètes*, Cambridge University Press, 2009, 354–398. https://www.cambridge.org/core/product/identifier/CBO9780511702709A018/type/book_part.

[38] Chan, H., B. Narayanan, M. J. Cherukara, F. G. Sen, K. Sasikumar, S. K. Gray, M. K. Chan, and S. K. Sankaranarayanan. "Machine Learning Classical Interatomic Potentials for Molecular Dynamics from First-Principles Training Data." *Journal of Physical Chemistry C* 123, 12: (2019) 6941–6957. https://pubs.acs.org/doi/full/10.1021/acs.jpcc.8b09917.

[39] Cherukara, M. J., B. Narayanan, A. Kinaci, K. Sasikumar, S. K. Gray, M. K. Chan, and S. K. Sankaranarayanan. "Ab Initio-Based Bond Order Potential to Investigate Low Thermal Conductivity of Stanene Nanostructures." *Journal of Physical Chemistry Letters* 7, 19: (2016) 3752–3759. https://pubs.acs.org/doi/full/10.1021/acs.jpclett.6b01562.

[40] Čižek, J., and J. Paldus. "Correlation problems in atomic and molecular systems III. Rederivation of the coupled-pair many-electron theory using the traditional quantum chemical methodst." *International Journal of Quantum Chemistry* 5, 4: (1971) 359–379. https://doi.org/10.1002/qua.560050402.

[41] Čížek, J. "On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods." *The Journal of Chemical Physics* 45, 11: (1966) 4256–4266. http://aip.scitation.org/doi/10.1063/1.1727484.

[42] ———. "On the Use of the Cluster Expansion and the Technique of Diagrams in Calculations of Correlation Effects in Atoms and Molecules." *Advances in Chemical Physics* 14: (1969) 35–89. https://onlinelibrary.wiley.com/doi/10.1002/9780470143599.ch2.

[43] Coester, F. "Bound states of a many-particle system." *Nuclear Physics* 7: (1958) 421–424.

[44] Coester, F., and H. Kümmel. "Short-range correlations in nuclear wave functions." *Nuclear Physics* 17: (1960) 477–485.

[45] Cromer, D. T., R. L. Mills, D. Schiferi, and L. A. Schwalbe. "The structure of N2 at 49 kbar and 299 K." *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* 37, 1: (1981) 8–11. https://scripts.iucr.org/cgi-bin/paper?S0567740881002070.

[46] Cyrot-Lackmann, F. "On the calculation of surface tension in transition metals." *Surface Science* 15, 3: (1969) 535–548.

[47] Daw, M. S., and M. I. Baskes. "Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals." *Physical Review Letters* 50, 17: (1983) 1285–1288.

[48] ———. "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals." *Physical Review B* 29, 12: (1984) 6443–6453.

[49] De Wette, F. W. "On the theory of transitions in some molecular crystals II." *Physica* 22, 6-12: (1956) 644–646.

[50] Deringer, V. L., M. A. Caro, and G. Csányi. "Machine Learning Interatomic Potentials as Emerging Tools for Materials Science." *Advanced Materials* 31, 46.

[51] Dirac, P. A. M. "On the theory of quantum mechanics." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 112, 762: (1926) 661–677. https://royalsocietypublishing.org/doi/10.1098/rspa.1926.0133.

[52] ———. "Note on Exchange Phenomena in the Thomas Atom." *Mathematical Proceedings of the Cambridge Philosophical Society* 26, 3: (1930) 376–385. https://www.cambridge.org/core/product/identifier/S0305004100016108/type/journal_article.

[53] Drautz, R. "Atomic cluster expansion for accurate and transferable interatomic potentials." *Physical Review B* 99, 1: (2019) 014,104.

[54] Dusson, G., M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner. "Atomic cluster expansion: Completeness, efficiency and stability." *Journal of Computational Physics* 454: (2022) 110,946.

[55] Elstner, M., D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert. "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties." *Physical Review B* 58, 11: (1998) 7260–7268.

[56] Engel, Y., S. Mannor, and R. Meir. "The kernel recursive least-squares algorithm." *IEEE Transactions on Signal Processing* 52, 8: (2004) 2275–2285.

[57] Eremets, M. I., A. G. Gavriliuk, N. R. Serebryanaya, I. A. Trojan, D. A. Dzivenko, R. Boehler, H. K. Mao, and R. J. Hemley. "Structural transformation of molecular nitrogen to a single-bonded atomic state at high pressures." *The Journal of Chemical Physics* 121, 22: (2004) 11,296.

[58] Eremets, M. I., R. J. Hemley, H.-k. Mao, and E. Gregoryanz. "Semiconducting non-molecular nitrogen up to 240 GPa and its low-pressure stability." *Nature* 411, 6834: (2001) 170–174. `http://www.nature.com/articles/35075531`.

[59] Eucken, A. "Über das thermische Verhalten einiger komprimierter und kondensierter Gase bei tiefen Temperaturen." *Verhandl. deut. physik. Ges* 18: (1916) 4–17.

[60] Fermi, E. "Un metodo statistico per la determinazione di alcune priorieta dell'atome." *Rend. Accad. Naz. Lincei* 6, 602-607: (1927) 32.

[61] Feynman, R. P. "Forces in Molecules." *Physical Review* 56, 4: (1939) 340–343. `https://link.aps.org/doi/10.1103/PhysRev.56.340`.

[62] Feynman, R. P., R. B. Leighton, M. Sands, and E. M. Hafner. *The Feynman Lectures on Physics; Vol. I*, volume 33. AAPT, 1965.

[63] Finney, A. R., and P. M. Rodger. "Applying the Z method to estimate temperatures of melting in structure II clathrate hydrates." *Physical Chemistry Chemical Physics* 13, 44: (2011) 19,979.

[64] Finnis, M. W., and J. E. Sinclair. "A simple empirical ¡i¿N¡/i¿ -body potential for transition metals." *Philosophical Magazine A* 50, 1: (1984) 45–55. `https://www.tandfonline.com/doi/full/10.1080/01418618408244210`.

[65] Frenkel, Y. I. *Kinetic Theory of Liquids*. (Oxford University Pres, 1946.

[66] Gamma, E., Richard Helm (Computer scientist), R. E. Johnson, and J. Vlissides. *Design patterns: Elements of reusable object-oriented software*. Addison-Wesley Professional, 1995.

[67] Gao, X., and L.-M. Duan. "Efficient representation of quantum many-body states with deep neural networks." *Nature Communications* 8, 1: (2017) 662.

[68] Gassner, H., M. Probst, A. Lauenstein, and K. Hermansson. "Representation of Intermolecular Potential Functions by Neural Networks." *Journal of Physical Chemistry A* 102, 24: (1998) 4596–4605. `https://pubs.acs.org/doi/abs/10.1021/jp972209d`.

[69] Gastegger, M., L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand. "wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials." *The Journal of Chemical Physics* 148, 24: (2018) 241,709. `https://aip.scitation.org/doi/abs/10.1063/1.5019667`.

[70] Gasteiger, J., and J. Zupan. "Neural Networks in Chemistry." *Angewandte Chemie International Edition in English* 32, 4: (1993) 503–527.

[71] Giannozzi, P., S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials." *Journal of Physics: Condensed Matter* 21, 39: (2009) 395,502.

[72] van der Giessen, E., P. A. Schultz, N. Bertin, V. V. Bulatov, W. Cai, G. Csányi, S. M. Foiles, M. G. D. Geers, C. González, M. Hütter, W. K. Kim, D. M. Kochmann, J. LLorca, A. E. Mattsson, J. Rottler, A. Shluger, R. B. Sills, I. Steinbach, A. Strachan, and E. B. Tadmor. "Roadmap on multiscale materials modeling." *Modelling and Simulation in Materials Science and Engineering* 28, 4: (2020) 043,001.

[73] Goldfeld, S. M., R. E. Quandt, and H. F. Trotter. "Maximization by Quadratic Hill-Climbing." *Econometrica* 34, 3: (1966) 541.

[74] Goncharov, A. F., E. Gregoryanz, H.-k. Mao, Z. Liu, and R. J. Hemley. "Optical Evidence for a Nonmolecular Phase of Nitrogen above 150 GPa." *Physical Review Letters* 85, 6: (2000) 1262–1265.

[75] Gregoryanz, E., A. F. Goncharov, C. Sanloup, M. Somayazulu, H.-k. Mao, and R. J. Hemley. "High P-T transformations of nitrogen to 170GPa." *The Journal of Chemical Physics* 126, 18: (2007) 184,505. `http://aip.scitation.org/doi/10.1063/1.2723069`.

[76] Gubaev, K. "Machine-Learning Interatomic Potentials for Multicomponent Alloys." *The Journal of Chemical Physics* 148, 24: (2018) 241,727.

[77] Gubaev, K., E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev. "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials." *Computational Materials Science* 156: (2019) 148–156.

[78] Gubaev, K., E. V. Podryabinkin, and A. V. Shapeev. "Machine learning of molecular properties: Locality and active learning." *The Journal of Chemical Physics* 148, 24: (2018) 241,727. https://aip.scitation.org/doi/abs/10.1063/1.5005095.

[79] Guennebaud, G., B. Jacob, and others. "Eigen v3." http://eigen.tuxfamily.org, 2010.

[80] Hale, L. M., Z. T. Trautt, and C. A. Becker. "Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants." *Modelling and Simulation in Materials Science and Engineering* 26, 5: (2018) 055,003.

[81] Hanfland, M., M. Lorenzen, C. Wassilew-Real, and F. Zontone. "Structures of Molecular Nitrogen at High Pressures." *Rev. High Pressure Sci. Technol* 7: (1998) 787–789.

[82] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York, 2009. http://link.springer.com/10.1007/978-0-387-84858-7.

[83] Hatfield, P. W., J. A. Gaffney, G. J. Anderson, S. Ali, L. Antonelli, S. Başeğmez du Pree, J. Citrin, M. Fajardo, P. Knapp, B. Kettle, B. Kustowski, M. J. MacDonald, D. Mariscal, M. E. Martin, T. Nagayama, C. A. J. Palmer, J. L. Peterson, S. Rose, J. J. Ruby, C. Shneider, M. J. V. Streeter, W. Trickey, and B. Williams. "The data-driven future of high-energy-density physics." *Nature* 593, 7859: (2021) 351–361.

[84] Heisenber, V. W., and i. Kopenhagen. "Mehrkörperproblem und Resonanz in der Quantenmechanik." *Zeitschrift für Physik 1926 38:6* 38, 6: (1926) 411–426. https://link.springer.com/article/10.1007/BF01397160.

[85] Hellmann, R. "Ab initio potential energy surface for the nitrogen molecule pair and thermophysical properties of nitrogen gas." *Molecular Physics* 111, 3: (2013) 387–401. http://www.tandfonline.com/doi/abs/10.1080/00268976.2012.726379.

[86] Hernandez, A., A. Balasubramanian, F. Yuan, S. A. Mason, and T. Mueller. "Fast, accurate, and transferable many-body interatomic potentials by symbolic regression." *npj Computational Materials 2019 5:1* 5, 1: (2019) 1–11. https://www.nature.com/articles/s41524-019-0249-1.

[87] Hernández, E., M. Gillan, and C. Goringe. "Basis functions for linear-scaling first-principles calculations." *Physical Review B - Condensed Matter and Materials Physics* 55, 20: (1997) 13,485–13,493.

[88] Hohenberg, P., and W. Kohn. "Inhomogeneous electron gas." *Physical Review* 136, 3B: (1964) B864. https://journals.aps.org/pr/abstract/10.1103/PhysRev.136.B864.

[89] Hoover, W. G. "Canonical dynamics: Equilibrium phase-space distributions." *Physical Review A* 31, 3: (1985) 1695. https://journals.aps.org/pra/abstract/10.1103/PhysRevA.31.1695.

[90] Hoover, W. G., and F. H. Ree. "Melting Transition and Communal Entropy for Hard Spheres." *The Journal of Chemical Physics* 49, 8: (2003) 3609. https://aip.scitation.org/doi/abs/10.1063/1.1670641.

[91] Horsfield, A., A. Bratkovsky, M. Fearn, D. Pettifor, and M. Aoki. "Bond-order potentials: Theory and implementation." *Physical Review B* 53, 19: (1996) 12,694. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.53.12694.

[92] Hutter, F., H. H. Hoos, and K. Leyton-Brown. "Sequential model-based optimization for general algorithm configuration." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6683 LNCS: (2011) 507–523. https://link.springer.com/chapter/10.1007/978-3-642-25566-3_40.

[93] Hutter, F., L. Kotthoff, and J. Vanschoren, editors. *Automated Machine Learning*. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019. http://link.springer.com/10.1007/978-3-030-05318-5.

[94] Inui, N. "Layered structure of Lennard-Jones particle systems confined in a step-shaped gap." *AIP Advances* 9, 7: (2019) 075,315. https://aip.scitation.org/doi/abs/10.1063/1.5096804.

[95] Jacobsen, R. T., R. B. Stewart, and M. Jahangiri. "Thermodynamic Properties of Nitrogen from the Freezing Line to 2000 K at Pressures to 1000 MPa." *Journal of Physical and Chemical Reference Data* 15, 2: (1986) 735–909. https://pubs.aip.org/aip/jpr/article/15/2/735/241330/Thermodynamic-Properties-of-Nitrogen-from-the.

[96] Jäger, B., R. Hellmann, and E. Bich. "State-of-the-art ab initio potential energy curve for the krypton atom pair and thermophysical properties of dilute krypton gas ARTICLES YOU MAY BE INTERESTED IN." *J. Chem. Phys* 144: (2016) 114,304. https://doi.org/10.1063/1.4943959.

[97] Jensen Frank. "Introduction to Computational Chemistry, 3rd Edition — Wiley." *WILEY* 660. https://www.wiley.com/en-us/Introduction+to+Computational+Chemistry%2C+3rd+Edition-p-9781118825990.

[98] Jones, J. E. "On the determination of molecular fields. —II. From the equation of state of a gas." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 106, 738: (1924) 463–477. https://royalsocietypublishing.org/doi/10.1098/rspa.1924.0082.

[99] ———. "On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 106, 738: (1924) 441–462. https://royalsocietypublishing.org/doi/10.1098/rspa.1924.0081.

[100] Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives. "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids." *Journal of the American Chemical Society* 118, 45: (1996) 11,225–11,236. https://pubs.acs.org/doi/abs/10.1021/ja9621760.

[101] Jorgensen, W. L., and J. Tirado-Rives. "The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin." *Journal of the American Chemical Society* 110, 6: (1988) 1657–1666. https://pubs.acs.org/doi/abs/10.1021/ja00214a001.

[102] Kennedy, J., and R. Eberhart. "Particle swarm optimization." In *Proceedings of ICNN'95 - International Conference on Neural Networks*. IEEE, 1995, volume 4, 1942–1948. http://ieeexplore.ieee.org/document/488968/.

[103] King, D. E. "Dlib-ml: A Machine Learning Toolkit." *Journal of Machine Learning Research* 10: (2009) 1755–1758.

[104] Kohn, W., and L. J. Sham. "Self-consistent equations including exchange and correlation effects." *Physical Review* 140, 4A: (1965) A1133. https://journals.aps.org/pr/abstract/10.1103/PhysRev.140.A1133.

[105] Kresse, G., and J. Hafner. "¡i¿Ab initio¡/i¿ molecular dynamics for liquid metals." *Physical Review B* 47, 1: (1993) 558–561.

[106] Larsson, H. R., A. C. Van Duin, and B. Hartke. "Global optimization of parameters in the reactive force field ReaxFF for SiOH." *Journal of Computational Chemistry* 34, 25: (2013) 2178–2189.

[107] Le, T., J. L. Doménech, M. Lepère, and H. Tran. "Molecular dynamic simulations of N2-broadened methane line shapes and comparison with experiments." *Journal of Chemical Physics* 146, 9. /aip/jcp/article/146/9/094305/76951/Molecular-dynamic-simulations-of-N2-broadened.

[108] Lee, C., W. Yang, and R. G. Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density."

Physical Review B 37, 2: (1988) 785–789. https://link.aps.org/doi/10.1103/PhysRevB.37.785.

[109] Lee, G., S. Bacon, I. Bush, L. Fortunato, D. Gavaghan, T. Lestang, C. Morton, M. Robinson, P. Rocca-Serra, S.-A. Sansone, and H. Webb. "Barely sufficient practices in scientific computing." *Patterns* 2, 2: (2021) 100,206.

[110] Leonhard, K., and U. K. Deiters. "Monte Carlo simulations of nitrogen using an ab initio potential." *https://doi.org/10.1080/00268970110118303* 100, 15: (2009) 2571–2585. https://www.tandfonline.com/doi/abs/10.1080/00268970110118303.

[111] LeSar, R., S. Ekberg, L. Jones, R. Mills, L. Schwalbe, and D. Schiferl. "Raman spectroscopy of solid nitrogen up to 374 kbar." *Solid State Communications* 32, 2: (1979) 131–134. https://linkinghub.elsevier.com/retrieve/pii/0038109879910731.

[112] Lessmann, S., R. Stahlbock, S. C. IC-AI, and u. 2005. "Optimizing hyperparameters of support vector machines by genetic algorithms." *sven-crone.info* http://sven-crone.info/papers/Lessmann,%20Stahlbock,%20Crone%20(2005)%20Optimizing%20Hyperparameters%20of%20Support%20Vector%20Machines%20by%20Genetic%20Algorithms%20ICAI05%20-%20ICA3442.pdf.

[113] Li, Y., H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Chan, S. K. Sankaranarayanan, B. R. Brooks, and B. Roux. "Machine Learning Force Field Parameters from Ab Initio Data." *Journal of Chemical Theory and Computation* 13, 9: (2017) 4492–4503. https://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00521.

[114] Lin, Y.-S., G. P. P. Pun, and Y. Mishin. "Development of a physically-informed neural network interatomic potential for tantalum." *Computational Materials Science* 205: (2022) 111,180. https://linkinghub.elsevier.com/retrieve/pii/S0927025621008338.

[115] Loach, C. H., and G. J. Ackland. "Stacking Characteristics of Close Packed Materials." *Physical Review Letters* 119, 20: (2017) 205,701. https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.119.205701.

[116] Lorenz, S., A. Groß, and M. Scheffler. "Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks." *Chemical Physics Letters* 395, 4-6: (2004) 210–215.

[117] Lorenzo, P. R., J. Nalepa, M. Kawulok, L. S. Ramos, and J. R. Pastor. "Particle swarm optimization for hyper-parameter selection in deep neural networks." In *Proceedings of the Genetic and Evolutionary Computation Conference*. New York, NY, USA: ACM, 2017, volume 8, 481–488. https://dl.acm.org/doi/10.1145/3071178.3071208.

[118] Luo, G. "A review of automatic selection methods for machine learning algorithms and hyper-parameter values." *Network Modeling Analysis in Health Informatics and Bioinformatics* 5, 1: (2016) 18. http://link.springer.com/10.1007/s13721-016-0125-6.

[119] Lysogorskiy, Y., C. v. d. Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, and R. Drautz. "Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon." *npj Computational Materials 2021 7:1* 7, 1: (2021) 1–12. https://www.nature.com/articles/s41524-021-00559-9.

[120] Malherbe, C., and N. Vayatis. "Global Optimization of Lipschitz Functions." In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. JMLR.org, 2017, ICML'17, 2314–2323.

[121] Marukatat, S. "Kernel matrix decomposition via empirical kernel map." *Pattern Recognition Letters* 77: (2016) 50–57.

[122] McCulloch, W. S., and W. Pitts. "A logical calculus of the ideas immanent in nervous activity." *The Bulletin of Mathematical Biophysics* 5, 4: (1943) 115–133.

[123] Mendelev, M. I., S. Han, D. J. Srolovitz, G. J. Ackland, D. Y. Sun, and M. Asta. "Development of new interatomic potentials appropriate for crystalline and liquid iron." *Philosophical Magazine* 83, 35: (2003) 3977–3994.

[124] Mendelev, M. I., T. L. Underwood, and G. J. Ackland. "Development of an interatomic potential for the simulation of defects, plasticity, and phase transformations in titanium." *The Journal of Chemical Physics* 145, 15: (2016) 154,102. http://aip.scitation.org/doi/10.1063/1.4964654.

[125] Miehlich, B., A. Savin, H. Stoll, and H. Preuss. "Results obtained with the correlation energy density functionals of becke and Lee, Yang and Parr." *Chemical Physics Letters* 157, 3: (1989) 200–206. https://linkinghub.elsevier.com/retrieve/pii/0009261489872343.

[126] Mills, R. L., B. Olinger, and D. T. Cromer. "Structures and phase diagrams of N ¡sub¿2¡/sub¿ and CO to 13 GPa by x-ray diffraction." *The Journal of Chemical Physics* 84, 5: (1986) 2837–2845. http://aip.scitation.org/doi/10.1063/1.450310.

[127] Mishin, Y. "Machine-learning interatomic potentials for materials science." *Acta Materialia* 214: (2021) 116,980.

[128] Moriarty, J. A. "Density-functional formulation of the generalized pseudopotential theory." *Physical Review B* 16, 6: (1977) 2537. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.16.2537.

[129] ———. "Density-functional formulation of the generalized pseudopotential theory. III. Transition-metal interatomic potentials." *Physical Review B* 38, 5: (1988) 3199–3231.

[130] Morse, P. M. "Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels." *Physical Review* 34, 1: (1929) 57. https://journals.aps.org/pr/abstract/10.1103/PhysRev.34.57.

[131] Murphy, K. P. *Machine Learning*. Adaptive Computation and Machine Learning series. London, England: MIT Press, 2012.

[132] Nichol, A., and G. J. Ackland. "Property trends in simple metals: An empirical potential approach." *Physical Review B* 93, 18: (2016) 184,101.

[133] Nosé, S. "A unified formulation of the constant temperature molecular dynamics methods." *The Journal of Chemical Physics* 81, 1: (1984) 511–519. http://aip.scitation.org/doi/10.1063/1.447334.

[134] Novikov, I. S., K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev. "The MLIP package: moment tensor potentials with MPI and active learning." *Machine Learning: Science and Technology* 2, 2: (2020) 025,002. https://iopscience.iop.org/article/10.1088/2632-2153/abc9fehttps://iopscience.iop.org/article/10.1088/2632-2153/abc9fe/meta.

[135] Olijnyk, H. "High pressure x-ray diffraction studies on solid N ¡sub¿2¡/sub¿ up to 43.9 GPa." *The Journal of Chemical Physics* 93, 12: (1990) 8968–8972.

[136] Pahari, P., and S. Chaturvedi. "Determination of best-fit potential parameters for a reactive force field using a genetic algorithm." *Journal of Molecular Modeling* 18, 3: (2012) 1049–1061. https://link.springer.com/article/10.1007/s00894-011-1124-2.

[137] Parrinello, M., and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method." *Journal of Applied Physics* 52, 12: (1998) 7182. https://aip.scitation.org/doi/abs/10.1063/1.328693.

[138] Perdew, J. P., K. Burke, and M. Ernzerhof. "Generalized Gradient Approximation Made Simple." *Physical Review Letters* 77, 18: (1996) 3865–3868. https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[139] Perdew, J. P., J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais. "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation." *Physical Review B* 46, 11: (1992) 6671. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.46.6671.

[140] Piyavskii, S. A. "An algorithm for finding the absolute extremum of a function." *USSR Computational Mathematics and Mathematical Physics* 12, 4: (1972) 57–67.

[141] Pozdnyakov, S. N., M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti. "Incompleteness of Atomic Structure Representations." *Physical Review Letters* 125, 16: (2020) 166,001.

[142] Proctor, J. E., C. G. Pruteanu, I. Morrison, I. F. Crowe, and J. S. Loveday. "Transition from Gas-like to Liquid-like Behavior in Supercritical N2." *Journal of Physical Chemistry Letters* 10, 21: (2019) 6584–6589. https://pubs.acs.org/doi/abs/10.1021/acs.jpclett.9b02358.

[143] Pruteanu, C. G., M. N. Bannerman, M. Kirsz, L. Lue, and G. J. Ackland. "From Atoms to Colloids: Does the Frenkel Line Exist in Discontinuous Potentials?" *ACS Omega* https://pubs.acs.org/doi/full/10.1021/acsomega.2c08056.

[144] Pruteanu, C. G., M. Kirsz, and G. J. Ackland. "Frenkel Line in Nitrogen Terminates at the Triple Point." *The Journal of Physical Chemistry Letters* 12, 47: (2021) 11,609–11,615. https://pubs.acs.org/doi/10.1021/acs.jpclett.1c03206.

[145] Pruteanu, C. G., J. E. Proctor, O. L. Alderman, and J. S. Loveday. "Structural Markers of the Frenkel Line in the Proximity of Widom Lines." *Journal of Physical Chemistry B* 125, 31: (2021) 8902–8906. https://pubs.acs.org/doi/abs/10.1021/acs.jpcb.1c04690.

[146] Pun, G. P. P., R. Batra, R. Ramprasad, and Y. Mishin. "Physically informed artificial neural networks for atomistic modeling of materials." *Nature Communications* 10, 1: (2019) 2339.

[147] Rappé, A. K., C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. "UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations." *Journal of the American Chemical Society* 114, 25: (1992) 10,024–10,035. https://pubs.acs.org/doi/abs/10.1021/ja00051a040.

[148] Rasmussen, C. E., and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[149] Ravelo, R., T. C. Germann, O. Guerrero, Q. An, and B. L. Holian. "Shock-induced plasticity in tantalum single crystals: Interatomic potentials and large-scale molecular-dynamics simulations." *Physical Review B - Condensed Matter and Materials Physics* 88, 13: (2013) 134,101.

[150] Robinson, D. W. "An experimental determination of the melting curves of argon and nitrogen into the 10000 atm region." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 225, 1162: (1954) 393–405. https://royalsocietypublishing.org/doi/10.1098/rspa.1954.0211.

[151] Rowe, P., V. L. Deringer, P. Gasparotto, G. Csányi, and A. Michaelides. "An accurate and transferable machine learning potential for carbon." *The Journal of Chemical Physics* 153, 3: (2020) 034,702.

[152] Ryckaert, J.-P., G. Ciccotti, and H. J. C. Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." *Journal of Computational Physics* 23, 3: (1977) 327–341. https://www.sciencedirect.com/science/article/pii/0021999177900985.

[153] Schiferl, D., S. Buchsbaum, and R. L. Mills. "Phase transitions in nitrogen observed by Raman spectroscopy from 0.4 to 27.4 GPa at 15 K." *The Journal of Physical Chemistry* 89, 11: (1985) 2324–2330. https://pubs.acs.org/doi/abs/10.1021/j100257a036.

[154] Schölkopf, B., S. Mika, C. J. Burges, P. Knirsch, K. R. Müller, G. Rätsch, and A. J. Smola. "Input space versus feature space in kernel-based methods." *IEEE Transactions on Neural Networks* 10, 5: (1999) 1000–1017.

[155] Schuch, A. F., and R. L. Mills. "Crystal Structures of the Three Modifications of Nitrogen 14 and Nitrogen 15 at High Pressure." *The Journal of Chemical Physics* 52, 12: (1970) 6000–6008. https://pubs.aip.org/aip/jcp/article/52/12/6000-6008/773397.

[156] Schumaker, L. *Spline Functions: Basic Theory*. Cambridge mathematical library. Cambridge: Cambridge University Press, 2007.

[157] Schwerdtfeger, P. "The Pseudopotential Approximation in Electronic Structure Theory." *ChemPhysChem* 12, 17: (2011) 3143–3155. https://onlinelibrary.wiley.com/doi/full/10.1002/cphc.201100387https://onlinelibrary.wiley.com/doi/abs/10.1002/cphc.201100387https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.201100387.

[158] Scurlock, R. G. *History and origins of cryogenics*. Oxford [England] ; New York : Oxford University Press, 1992.

[159] Seeger, M. "Gaussian processes for machine learning." *International journal of neural systems* 14, 2: (2004) 69–106.

[160] Segall, M. D., P. J. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne. "First-principles simulation: Ideas, illustrations and the CASTEP code.", 2002. https://iopscience.iop.org/article/10.1088/0953-8984/14/11/301/meta.

[161] Sen, F. G., A. Kinaci, B. Narayanan, S. K. Gray, M. J. Davis, S. K. Sankaranarayanan, and M. K. Chan. "Towards accurate prediction of catalytic activity in IrO2 nanoclusters via first principles-based variable charge force field." *Journal of Materials Chemistry A* 3, 37: (2015)

18,970–18,982. https://pubs.rsc.org/en/content/articlehtml/2015/ta/c5ta04678ehttps://pubs.rsc.org/en/content/articlelanding/2015/ta/c5ta04678e.

[162] Shapeev, A. V. "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials." *Multiscale Modeling & Simulation* 14, 3: (2016) 1153–1173. http://epubs.siam.org/doi/10.1137/15M1054183.

[163] Shi, Y., and R. C. Eberhart. "Parameter selection in particle swarm optimization." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1447: (1998) 591–600.

[164] Shubert, B. O. "A Sequential Method Seeking the Global Maximum of a Function." *https://doi.org/10.1137/0709036* 9, 3: (2006) 379–388. https://epubs.siam.org/doi/10.1137/0709036.

[165] Simeoni, G. G., T. Bryk, F. A. Gorelli, M. Krisch, G. Ruocco, M. Santoro, and T. Scopigno. "The Widom line as the crossover between liquid-like and gas-like behaviour in supercritical fluids." *Nature Physics 2010 6:7* 6, 7: (2010) 503–507. https://www.nature.com/articles/nphys1683.

[166] Slater, J. C. "The Theory of Complex Spectra." *Physical Review* 34, 10: (1929) 1293. https://journals.aps.org/pr/abstract/10.1103/PhysRev.34.1293.

[167] Slater, J. C., and G. F. Koster. "Simplified LCAO Method for the Periodic Potential Problem." *Physical Review* 94, 6: (1954) 1498. https://journals.aps.org/pr/abstract/10.1103/PhysRev.94.1498.

[168] Slater, J. C. "Atomic Shielding Constants." *Physical Review* 36, 1: (1930) 57–64. https://link.aps.org/doi/10.1103/PhysRev.36.57.

[169] Smith, D., M. A. Hakeem, P. Parisiades, H. E. Maynard-Casely, D. Foster, D. Eden, D. J. Bull, A. R. Marshall, A. M. Adawi, R. Howie, A. Sapelkin, V. V. Brazhkin, and J. E. Proctor. "Crossover between liquidlike and gaslike behavior in C H4 at 400 K." *Physical Review E* 96, 5: (2017) 052,113. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.96.052113.

[170] Snoek, J., H. Larochelle, and R. P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms." In *Advances in Neural Information Processing Systems*, edited by F Pereira, C J Burges, L Bottou, and K Q Weinberger. Curran Associates, Inc., 2012, volume 25. https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

[171] Sorensen, D. C. "Newton's Method with a Model Trust Region Modification." *SIAM Journal on Numerical Analysis* 19, 2: (1982) 409–426. http://epubs.siam.org/doi/10.1137/0719026.

[172] Span, R., E. W. Lemmon, R. T. Jacobsen, W. Wagner, and A. Yokozeki. "A Reference Equation of State for the Thermodynamic Properties of Nitrogen for Temperatures from 63.151 to 1000 K and Pressures to 2200 MPa." *Journal of Physical and Chemical Reference Data* 29, 6: (2000) 1361–1433. http://aip.scitation.org/doi/10.1063/1.1349047.

[173] Sparks, E. R., A. Talwalkar, D. Haas, M. J. Franklin, M. I. Jordan, and T. Kraska. "Automating model search for large scale machine learning." In *Proceedings of the Sixth ACM Symposium on Cloud Computing*. New York, NY, USA: ACM, 2015, 368–380. https://dl.acm.org/doi/10.1145/2806777.2806945.

[174] Stillinger, F. H., and T. A. Weber. "Computer simulation of local order in condensed phases of silicon." *Physical Review B* 31, 8: (1985) 5262. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.31.5262.

[175] Stinton, G. W., I. Loa, L. F. Lundegaard, and M. I. McMahon. "The crystal structures of δ and δ* nitrogen." *The Journal of Chemical Physics* 131, 10: (2009) 104,511. http://scitation.aip.org/content/aip/journal/jcp/131/10/10.1063/1.3204074.

[176] Strąk, P., and S. Krukowski. "Molecular nitrogen-N2 properties: The intermolecular potential and the equation of state." *The Journal of Chemical Physics* 126, 19: (2007) 194,501. https://pubs.aip.org/aip/jcp/article/929612.

[177] Stroppa, A., and G. Kresse. "The shortcomings of semi-local and hybrid functionals: what we can learn from surface science studies." *New Journal of Physics* 10, 6: (2008) 063,020. https://iopscience.iop.org/article/10.1088/1367-2630/10/6/063020.

[178] Sumpter, B. G., and D. W. Noid. "Neural networks and graph theory as computational tools for predicting polymer properties." *Macromolecular Theory and Simulations* 3, 2: (1994) 363–378.

[179] Swenson, C. A. "New Modification of Solid Nitrogen." *The Journal of Chemical Physics* 23, 10: (1955) 1963–1964.

[180] Swope, W. C., H. C. Andersen, P. H. Berens, and K. R. Wilson. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters." *The Journal of Chemical Physics* 76, 1: (1998) 637. https://aip.scitation.org/doi/abs/10.1063/1.442716.

[181] Tadmor, E. B., R. S. Elliott, J. P. Sethna, R. E. Miller, and C. A. Becker. "The potential of atomistic simulations and the knowledgebase of interatomic models." *JOM* 63, 7: (2011) 17–17.

[182] Tadmor, E. B., and R. E. Miller. "Modeling materials: Continuum, atomistic and multiscale techniques." *Modeling Materials: Continuum, Atomistic and Multiscale Techniques* 9780521856980: (2011) 1–759. https://www-cambridge-org.ezproxy.is.ed.ac.uk/core/books/modeling-materials/7CC4027C34755637D8F641A0C8C26835.

[183] Takahashi, A., A. Seko, and I. Tanaka. "Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: Application to elemental titanium." *Physical Review Materials* 1, 6: (2017) 063,801.

[184] Tersoff, J. "Empirical interatomic potential for silicon with improved elastic properties." *Physical Review B* 38, 14: (1988) 9902–9905.

[185] ———. "New empirical approach for the structure and energy of covalent systems." *Physical Review B* 37, 12: (1988) 6991. https://journals.aps.org/prb/abstract/10.1103/PhysRevB.37.6991.

[186] ———. "Modeling solid-state chemistry: Interatomic potentials for multicomponent systems." *Physical Review B* 39, 8: (1989) 5566–5568.

[187] Thomas, L. H. "The calculation of atomic fields." *Mathematical Proceedings of the Cambridge Philosophical Society* 23, 5: (1927) 542–548. https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/abs/calculation-of-atomic-fields/ADCA3D21D0FACD7077B5FDBB7F3B3F3A.

[188] Thompson, A. P., L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker. "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials." *Journal of Computational Physics* 285: (2015) 316–330.

[189] Thompson, A. P., H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. Michael Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales." *Computer Physics Communications* 271: (2022) 108,171.

[190] Thompson, A. P., S. J. Plimpton, and W. Mattson. "General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions." *The Journal of Chemical Physics* 131, 15: (2009) 154,107. https://aip.scitation.org/doi/abs/10.1063/1.3245303.

[191] Thomsen, J., and B. Meyer. "Pattern recognition of the 1H NMR spectra of sugar alditols using a neural network." *Journal of Magnetic Resonance (1969)* 84, 1: (1989) 212–217.

[192] Tonkov, E. Y., and E. Ponyatovsky. *Phase Transformations of Elements Under High Pressure*. CRC Press, 2018.

[193] Truesdell, C., and W. Noll. "The Non-Linear Field Theories of Mechanics." *The Non-Linear Field Theories of Mechanics* .

[194] TURING, A. M. "I.—COMPUTING MACHINERY AND INTELLIGENCE." *Mind* LIX, 236: (1950) 433–460.

[195] Unruh, D., R. V. Meidanshahi, S. M. Goodnick, G. Csányi, and G. T. Zimányi. "Gaussian approximation potential for amorphous Si : H." *Physical Review Materials* 6, 6: (2022) 065,603.

[196] Van Der Oord, C., G. Dusson, G. Csányi, and C. Ortner. "Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials." *Machine Learning: Science and Technology* 1, 1: (2020) 015,004. https://iopscience.iop.org/article/10.1088/2632-2153/ab527chttps://iopscience.iop.org/article/10.1088/2632-2153/ab527c/meta.

[197] Venables, J. A., and C. A. English. "Electron diffraction and the structure of α-N2." *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* 30, 4: (1974) 929–935. https://scripts.iucr.org/cgi-bin/paper?S0567740874004067.

[198] Vikhar, P. A. "Evolutionary algorithms: A critical review and its future prospects." *Proceedings - International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016* 261–265.

[199] Vos, W. L., and J. A. Schouten. "Improved phase diagram of nitrogen up to 85 kbar." *The Journal of Chemical Physics* 91, 10: (1989) 6302–6305. http://aip.scitation.org/doi/10.1063/1.457397.

[200] Wang, H., L. Zhang, J. Han, and W. E. "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics." *Computer Physics Communications* 228: (2018) 178–184.

[201] Wen, M., Y. Afshar, R. S. Elliott, and E. B. Tadmor. "KLIFF: A framework to develop physics-based and machine learning interatomic potentials." *Computer Physics Communications* 272: (2022) 108,218.

[202] Whitley, D. "A genetic algorithm tutorial." *Statistics and Computing* 4, 2: (1994) 65–85. https://link.springer.com/article/10.1007/BF00175354.

[203] Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson. "Best Practices for Scientific Computing." *PLoS Biology* 12, 1: (2014) e1001,745.

[204] Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. "Good enough practices in scientific computing." *PLOS Computational Biology* 13, 6: (2017) e1005,510.

[205] Yang, J. X., H. L. Zhao, H. R. Gong, M. Song, and Q. Q. Ren. "Proposed mechanism of HCP to FCC phase transition in titianium through first principles calculation and experiments." *Scientific Reports* 8, 1: (2018) 1992. https://www.nature.com/articles/s41598-018-20257-9.

[206] Yang, L., and A. Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice." *Neurocomputing* 415: (2020) 295–316. https://linkinghub.elsevier.com/retrieve/pii/S0925231220311693.

[207] Yanxon, H., D. Zagaceta, B. Tang, D. S. Matteson, and Q. Zhu. "PyXtal_FF: a python library for automated force field generation OPEN ACCESS RECEIVED PyXtal_FF: a python library for automated force field generation." *Mach. Learn.: Sci. Technol* 2: (2021) 27,001. https://doi.org/10.1088/2632-2153/abc940.

[208] Young, D. A., C.-S. Zha, R. Boehler, J. Yen, M. Nicol, A. S. Zinn, D. Schiferl, S. Kinkead, R. C. Hanson, and D. A. Pinnick. "Diatomic melting curves to very high pressure." *Physical Review B* 35, 10: (1987) 5353–5356. https://link.aps.org/doi/10.1103/PhysRevB.35.5353.

[209] Zhang, I. Y., and A. Grüneis. "Coupled cluster theory in materials science." *Frontiers in Materials* 6: (2019) 123.

[210] Zhang, L., J. Han, H. Wang, R. Car, and E. Weinan. "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics." *Physical Review Letters* 120, 14: (2018) 143,001. https://link.aps.org/doi/10.1103/PhysRevLett.120.143001https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.143001.

[211] Zhang, Y., C. Hu, and B. Jiang. "Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation." *Journal of Physical Chemistry Letters* 10, 17: (2019) 4962–4967.

[212] Zhao, J., X. Li, C. Shum, and J. McPhee. "A Review of physics-based and data-driven models for real-time control of polymer electrolyte membrane fuel cells." *Energy and AI* 6: (2021) 100,114.

[213] Ziegler, J. F., and J. P. Biersack. "The Stopping and Range of Ions in Matter." *Treatise on Heavy-Ion Science* 93–129. https://link.springer.com/chapter/10.1007/978-1-4615-8103-1_3.

[214] Zöller, M., M. H. J. o. a. i. research, and u. 2021. "Benchmark and survey of automated machine learning frameworks." *jair.org* 70: (2021) 409–472. https://www.jair.org/index.php/jair/article/view/11854.

[215] Zong, H., G. Pilania, X. Ding, G. J. Ackland, and T. Lookman. "Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning." *npj Computational Materials* 4, 1: (2018) 48. http://www.nature.com/articles/s41524-018-0103-xhttps://www.nature.com/articles/s41524-018-0103-x.

[216] Zuo, Y., C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong. "Performance and Cost Assessment of Machine Learning Interatomic Potentials." *Journal of Physical Chemistry A* 124, 4: (2020) 731–745. http://arxiv.org/abs/1906.08888.