Sensitivity Analysis of Material Property Predictions in Molecular Dynamics with Interatomic Potentials under Uncertainty

by

Tomas Mawyin

A thesis submitted in conformity with the requirements for the degree of Master’s of Applied Science and Engineering Graduate Department of Aerospace Science and Engineering University of Toronto

© Copyright 2015 by Tomas Mawyin
Abstract

Sensitivity Analysis of Material Property Predictions in Molecular Dynamics with Interatomic Potentials under Uncertainty

Tomas Mawyin
Master’s of Applied Science and Engineering
Graduate Department of Aerospace Science and Engineering
University of Toronto
2015

Molecular Dynamics (MD) simulations complement experimental methods by offering a computationally fast and accurate framework for determining the properties of materials. However, there is uncertainty associated with MD simulations in part due to the intrinsic microstructure of the material, but also due to the interatomic potential used to describe the material. To evaluate the uncertainty associated with MD predictions, this work exploits the strong dependency of MD on the interatomic potentials. Using a methodology based on surrogate models, a sensitivity analysis of the properties of graphene and aluminium when uncertainty is introduced to the parameters of their respective potentials is presented. These surrogates are utilized to identify which parameters in the potentials greatly affect the output parameters of interest. This methodology can be used to evaluate the robustness of interatomic potentials and accelerate the solution of the inverse problem to determine the parameters of the potential from observables.
Dedication

This thesis is dedicated to my parents, for their endless love, support, and encouragement.

I would like to thank my supervisors Professor Prasanth. B. Nair, and Professor Chandra Veer Singh for all their support, expert advice and unfailing patience. I would also like to thank my friends who tirelessly listened to my ideas and offered encouragement and support when it was most needed.
## Contents

1 Introduction

1.1 Materials and Uncertainty Quantification ........................................ 1
1.2 Problem Statement .............................................................................. 4
1.3 Thesis Objectives .............................................................................. 5
1.4 Thesis Layout ..................................................................................... 5

2 Methodology

2.1 Molecular Dynamics Simulations ....................................................... 7
  2.1.1 Background and Applications ....................................................... 7
  2.1.2 General Theory ........................................................................... 8
    Equations of Motion ........................................................................... 9
    Temperature and Pressure Controls .................................................. 10
    Boundary Conditions ........................................................................ 11
  2.1.3 Limitations .................................................................................. 11
2.2 Uncertainty Propagation and Quantification ...................................... 13
  2.2.1 Background: Prediction of Errors ................................................. 13
  2.2.2 Monte Carlo ................................................................................ 14
  2.2.3 Sparse Quadrature ....................................................................... 16
  2.2.4 Latin Hypercube Sampling ........................................................ 18
2.3 Surrogate Model ................................................................................. 19
  2.3.1 Background and Applications ....................................................... 19
  2.3.2 Bayesian Gaussian Process Modelling ......................................... 20
  2.3.3 Radial Basis Function Surrogates ............................................... 22
  2.3.4 Reduced Order Modelling ........................................................... 24
  2.3.5 Surrogate Model Validation ......................................................... 25
  2.3.6 Sensitivity Analysis .................................................................... 26
3 2D Material System: Graphene

3.1 Graphene Topology ................................................. 28
  3.1.1 Defects in Graphene Topology ................................ 29
3.2 AIREBO Interatomic Potential ..................................... 30
3.3 Graphene Properties from Molecular Dynamics ..................... 31
  3.3.1 Elastic Constants ............................................. 31
  3.3.2 Energetic Properties and Defects .............................. 32
    Cohesive Energy .............................................. 32
    Defect Formation Energy ...................................... 32
    Grain Boundary Formation Energy .............................. 33
  3.3.3 Thermal Conductivity ........................................... 33
  3.3.4 Strain-Stress Relationship .................................... 34
3.4 Uncertainty Propagation and Quantification ........................ 35
  3.4.1 Uncertainty Propagation in LJ Potential ...................... 35
  3.4.2 Uncertainty Propagation in AIREBO Potential .................. 36
    Monte Carlo Analysis .......................................... 37
    Sparse Grid Analysis .......................................... 39
3.5 Surrogate Model Applied to Graphene ............................ 41
  3.5.1 Bayesian Gaussian Process Model ................................ 42
    Leave-one-out Validation ...................................... 42
    Standardized Cross-Validation Residuals ....................... 44
  3.5.2 Model Comparison: GPMs and RBFs ............................. 46
3.6 Sensitivity Analysis ............................................... 46
  3.6.1 Main Effects ................................................. 46
  3.6.2 Sensitivity Indexes .......................................... 48
  3.6.3 Interaction Effects ......................................... 52

4 3D Material Systems: Metals .......................................... 54

4.1 Aluminium Topology .............................................. 54
  4.1.1 Defects in FCC Structures ................................... 55
4.2 EAM Interatomic Potential ....................................... 57
4.3 Aluminium Properties from Molecular Dynamics .................... 59
  4.3.1 Elastic Constants ............................................ 59
  4.3.2 Cohesive Energy and Lattice Parameter ...................... 60
  4.3.3 Point Defects Energetic Properties .......................... 60
    Vacancy Formation Energy .................................... 60
Interstitial Formation Energy ........................................ 61
4.3.4 Energetic Properties of Planar Defects ....................... 61
   Stacking Fault Energy ............................................. 61
   Grain Boundary Energy .......................................... 62
   Free Surface Energy .............................................. 63
4.4 Uncertainty Propagation and Quantification ...................... 63
4.5 Surrogate Model Applied to Aluminium ......................... 67
   4.5.1 Bayesian Gaussian Process Model ........................... 67
       Leave-one-out Validation ..................................... 68
       Standardized Cross-Validation Residuals ................... 68
4.6 Sensitivity Analysis ............................................. 71
   4.6.1 Main Effects and Sensitivity Indexes ....................... 71
4.7 Reduced Order Modelling and RBF ............................... 72

5 Interatomic Potential from Surrogate Models .......................... 75
   5.1 Inverse Problem: Objectives .................................. 75
   5.2 Problem Statement ............................................ 76
       5.2.1 Procedure .............................................. 77
       5.2.2 Simulated Annealing Algorithm ......................... 78
   5.3 Study: Recovering Graphene’s Potential ....................... 79
   5.4 Study: Parameter Estimation with Noisy Observables .......... 81
   5.5 Concluding Remarks .......................................... 82

6 Conclusion and Future Work ........................................ 84
   6.1 Conclusions .................................................. 84
       6.1.1 Summary of Methodology ................................ 85
       6.1.2 Graphene ............................................... 86
       6.1.3 Aluminium .............................................. 86
       6.1.4 Inverse Problem ......................................... 87
   6.2 Future Work ................................................... 88
       6.2.1 MD Simulations .......................................... 88
       6.2.2 Surrogate Models ........................................ 89

Bibliography .......................................................... 90
## List of Tables

3.1 Reference values for graphene observables used in this thesis .......................... 35
4.1 Reference values for aluminium observables used in this thesis .......................... 63
5.1 Solution of optimization problem using four observables using different initial guesses (trials) .......................................................... 79
5.2 Comparison between MD simulation and GPM results when using recovered input parameters ......................................................... 80
5.3 Solution of optimization problem using five, six, and seven observables .......... 80
List of Figures

2.1 Display of periodic boundary conditions in a 2D system. Arrows represent the direction of motion of the particles. 12

2.2 Comparison between full tensor product and sparse tensor product to integrate a function of order 16 exactly. Full quadrature requires 289 points while the sparse quadrature requires only 65 [1]. 18

2.3 Comparison between a random sampling method and Latin Hypercube sampling. Using \( n = 10 \) points for two dimensions \( d = 2 \) ranging from \([0,1]\) 19

3.1 Graphene topology in the presence of different defects. The colour map defines the atomic energy in \([\text{eV}]\) for the graphene sheet. 34

3.2 Scatter plots for (a) \( C_{11} \), (b) \( C_{12} \) elastic constants, (c) cohesive energy, and (d) vacancy formation energy following a \( \pm 5\% \) variability in the LJ parameters of the AIREBO potential. 36

3.3 Statistical results of (a) \( C_{11} \), (b) \( C_{12} \) elastic constants, (c) cohesive energy, and (d) vacancy formation energy when variability is introduced to the AIREBO potential. 37

3.3 (cont’d) Statistical results of (e) divacancy formation energy, (f) bond rotation formation energy, (g) grain boundary formation energy, and (h) thermal conductivity when variability is introduced to the AIREBO potential. 38

3.4 Quantities of interest mean and variance when using sparse grid technique to sample parameter space. 40

3.5 (a) Graphene stress-strain curves using four different AIREBO potentials with a \( \pm 5\% \) variability. The solid black line represents the \( \sigma - \epsilon \) curve for pristine graphene using the nominal parameters of the interatomic potential. (b) MD tensile test simulation of pristine graphene sheet at breaking point. 41
3.6 Leave-one-out validation approach of the surrogate model for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacanty}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}}$. 43

3.7 Leave-p-out validation approach for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{divacanty}}$, and (d) $E_{\text{grain-boundary}}$ formation energies. Validation obtained using $p = 20$ training points. 44

3.8 Standardize cross-validation residuals for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacanty}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}}$. $\pm 3\sigma$ are marked with dotted lines. 45

3.9 GPM and RBF model comparison for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacanty}}$, (f) $E_{\text{bond-rotation}}$, and (g) $E_{\text{grain-boundary}}$ formation energies. A k-fold cross validation is used with $N = 250$ points for the training and testing sets. 47

3.10 Main effects for $C_{11}$ and $E_{\text{vacancy}}$ formation energy. Abscissa shows a normalized value for each input parameter in the AIREBO potential. Red dotted line signifies a $\pm 95\%$ confidence interval based on the standard error. 49

3.11 Main effects for $E_{\text{cohesive}}$ and $E_{\text{grain-boundary}}$ formation energy. Abscissa shows a normalized value for each input parameter in the AIREBO potential. Red dotted line signifies a $\pm 95\%$ confidence interval based on the standard error. 50

3.12 Pie charts showing the sensitivity of the parameters in the AIREBO potential for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacanty}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}}$. 51

3.13 Interaction effects from different input variables to describe the response of the $C_{11}$ elastic constant. 53

4.1 Face centred cubic or cubic close packed structure. 55

4.2 Individual functional forms of the aluminium’s EAM potential. 58

4.3 Vacancy and interstitial defects in a FCC structure. The colours represent the energy map of the system in [eV]. 61

4.4 Stacking fault and grain boundary formation of a FCC structure. The colours represent the energy map of the system in [eV]. 62

4.5 Comparison between original points and interpolation result for the embedding function in the EAM potential. The lower plot describes the interpolation error. 64
4.6 Boxplots showing the effects on the quantities of interest when uncertainty is introduced at different variabilities to the interpolation coefficients of the EAM potential. Range taken by the observables is shown for clarity. 65

4.6 (cont’d) Boxplots showing the effects on the quantities of interest when uncertainty is introduced at different variabilities to the interpolation coefficients of the EAM potential. Range taken by the observables is shown for clarity. 66

4.7 Leave-one-out validation approach of the surrogate model for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of $\pm5\%$ in the EAM potential for aluminium material system. 69

4.8 Standardized cross-validation residuals for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of $\pm5\%$. Dotted line represents the confident interval as $\pm3\sigma$. 70

4.9 Main effects for $E_{\text{grain-boundary}}$ formation energy when using a variability of $\pm5\%$. $X_i$ signifies the $i^{th}$ interpolation coefficient used as input parameter. 72

4.10 Sensitivity indices for (a) $C_{11}$ and (b) $E_{\text{grain-boundary}}$ formation energy. 73

4.11 Principal component analysis combined with RBF surrogate and error for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{vacancy}}$, (d) $E_{\text{interstitial}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of $\pm5\%$ in the EAM potential for aluminium material system. 74

5.1 Histograms displaying error percentage when noise is introduced to the experimental observables prior to the prediction of the input parameters. Dotted line represents the error mean of the distribution. 82
Chapter 1

Introduction

This chapter describes the background information that motivated the studies presented within this thesis. Relating the fields of materials science and statistics to incorporate techniques for uncertainty quantification, propagation, and surrogate models in the design or exploration of material systems is a relatively new idea. This chapter presents an overview of the current research in these fields as well as a layout of the problem statement and the objectives of this work.

1.1 Materials and Uncertainty Quantification

Some current innovations in material systems arise from the creation and modelling of nanomaterials. A nanomaterial can be described as a material containing particles where, for 50% of more of the particles in the number size distribution, one or more external dimensions is in the size range from 1 nm to 100 nm [2]. These materials are created to exhibit properties that can be used for innovative applications in the electronics, biomedical, energy or aerospace fields. In current applications, nanomaterials are being used in the construction of lightweight and high strength structures as well as in the creation of coatings to provide structures with higher thermal or corrosion resistance [3].

Nanomaterials have gained popularity because they can be tailored to provide properties that conventional macro-scale materials cannot provide. However, the creation of these “materials-by-design”, specifically at this scale, brings up other considerations such as the ability to experimentally test the materials to characterize their properties. Many methods that incorporate statistical analysis into the modelling of these materials have been proposed to provide a more realistic characterization of their material properties. Research has shown that at the nano-scale, the stochastic motion of atoms becomes significant, thus many material qualities follow statistical distributions [4].
When designing experiments to obtain mechanical properties of nanostructures such as nanowires or nanobelts, it has been observed that many errors and biases arise from the accuracy of the measurements, difficulty in manipulation, and the small size of the material [5, 6]. Material uncertainty, such as surface roughness or imperfect physical boundary conditions can also lead to large systematic errors and uncertainty in data quantification. At such small scales, any small error becomes detrimental in the quantification of the mechanical properties. One way to reduce biases that arise from the accuracy of the measurements is to use the variability encountered in the experimental results to incorporate a regression model to the physical experiment in order to capture initial and subsequent biases. This approach was used by Deng et al. [5] when measuring force-deflection curves of a nanostructure. In their work, a simply-supported beam model was used to obtain the elastic modulus; however, the experimentation showed systematic errors attributed to the size of the tip of the instrument performing the atomic force microscopy, the imperfections of the material, and the stability of the structure during measurement. The combination of the statistical and the physical model proved to be effective in determining the elastic modulus of ZnO nanobelts.

Identifying the proper distribution for the properties of nanomaterials tends to be a challenging task because the results arising from experimental data are distributed in a very narrow range. By using techniques such as the minimum (Akaike) information criterion, which are based on maximum likelihood, it is possible to objectively determine the distribution of the material property. This approach has been implemented by Lu [6] to determine the strength distribution of brittle and ductile materials as well as the size distribution of nano-grains which describes the hardness in the material.

Current research in nanomaterials goes far beyond exploring the challenges in experimental techniques; in fact, understanding the uncertainty in the material structure plays a fundamental role in defining the properties of the material. A key example is the distribution of grain size, shapes, and boundaries within the structure to determine the failure mechanisms and mechanical properties of the material. Zhu et al [7] demonstrated that a small grain size, up to 75 nm, provides the strength of these materials - represented by the Hall-Petch relationship. However, the study also points out that a decrease in the grain size is reflected in a decrease in the ductility of the material. By combining stochastic approaches based on probability distributions with deformation mechanisms, it is possible to accurately describe the failure properties of nanomaterials. By comparing grain sizes with the distribution of micro-cracks in the structure it was found that the stress and strain fields of the material change with these micro-cracks, which are then taken into consideration during the plastic deformation of the material [7].
Although the material structure defines its properties, it should be noted that the structure is greatly dependent on the fabrication process. As pointed out by Tengen [8], having two nanomaterials with the same mean grain size but produced through different techniques might display different mechanical properties. Moreover, nanomaterials with different mean grain sizes and dispersion might exhibit similar mechanical properties. The inclusion of stochastic techniques in the design and modelling of nanomaterials’ constituent and material properties is therefore imperative [8]. A model including the relationship between shape and size as well as the evolution for grain stress and strain to solve for the mechanical properties of the material has been proposed to characterize properties of nanocrystalline aluminum with successful results [8].

A fundamental piece in the development of materials at such scale is the use of atomistic models. Computational modelling using molecular dynamics (MD) seems to be an effective and efficient alternative to experimental methods and high fidelity \textit{ab initio} methods such as, the Hartre-Fock method and density functional theory. Using this methodology, it has been possible to study the size dependence, grain boundaries, and failure mechanisms, among other material properties with very promising results. In general, atomic level modelling can provide insights into the phase stability, structure and properties of crystalline defects, and into physical mechanisms of many processes ranging from atomic diffusion to interface migration [9].

Certainly, the process of linking uncertainty quantification with atomistic modelling of materials is fairly new. An approach to estimate the transferability of errors through the generation of ensembles of potentials to estimate error bars for the prediction of atomistic calculations has been developed [10]. This approach shows that uncertainty in modelling materials using MD is present due to the material definition as well as the methodology used to describe the atomic interactions. Quantifying and propagating parametric uncertainty in MD has also been the subject of a few recent studies. Rizzi et al [11] use simulations of water molecules combined with non-intrusive techniques such as polynomial chaos (PC) representation to investigate how uncertainty, in the most basic atomic interaction, can modify the overall properties of the material.

Finally, stochastic techniques have been applied to develop a high performance computing implementation to reduce time to solution and overall computational cost of uncertainty quantification in MD simulations [12]. These methodologies aim to identify which parameters from the MD interactions are crucial to the propagation of uncertainty.
1.2 Problem Statement

As described in the previous section, MD simulations have been used to determine how atomic or molecular interactions affect the macroscopic properties of different material systems [13, 14]. MD simulations complement experimental methods [15] and \textit{ab initio} calculations by offering a computationally fast and accurate framework for determining the properties of new or existing materials. This technique is becoming more popular in industry and academia for modelling materials at such small scale to gain fundamental insights into their performance and failure mechanisms [14, 16]. Besides the obvious limitations on length and time scales [9, 14, 17], another key limitation of existing approaches for using MD is the implicit assumption that the interatomic potential and the structure of the material are known precisely [17–19]. In practice, however, significant uncertainties may arise from the nature of the material, e.g., defects such as vacancies or grain boundaries [14], as well as the parameters used for specifying the interatomic potential [10, 12].

The main weakness of MD simulations is their strong dependency on interatomic potentials. A given limitation of these potentials is the prediction of atomic forces and energies without considering the quantum or electronic structure of the material [10, 17]. In general, interatomic potentials are created by fitting a curve to experimental observations and/or results from \textit{ab initio} calculations with the expectation that the potential will accurately make predictions of the properties of the material system as a whole [9, 12, 20]. This is because the curve fitting is geared to describe certain properties more accurately than others, that is, certain parameters are weighted heavier than others during the fitting procedure [18]. For instance, the Tersoff interatomic potential used to describe Silicon, adds higher weights to properties such as lattice parameter, cohesive energy, and bulk modulus during the fitting procedure. However, the Tersoff potential fails to accurately describe properties such as shear modulus or defect energies [21]. Additionally, to make simulations computationally less expensive, interatomic potentials are formulated in terms of the contribution of only a few neighbouring atoms which implies that there are approximation errors associated with the potential models [18].

Research studying the parametric uncertainty arising from the interatomic potential is currently limited. Although some work has been done for simple interatomic interactions, such as the Lennard Jones (LJ) potentials, most of the ground-breaking materials or even the most common materials used in daily applications use more complex interatomic potentials or interactions that usually depend on the combination of two or more potentials. This thesis aims to understand the sensitivity of the material properties given
uncertainty in the parameters of the interatomic potential by using stochastic techniques combined with MD simulations.

1.3 Thesis Objectives

To gain an understanding of the sensitivity of the materials properties based on the uncertainty in the parameters of the interatomic potentials when using MD, the following research objectives are defined:

1. Reproduce material topologies, including defects, to perform MD simulations in order to predict material properties.

2. Develop methodologies to propagate and quantify uncertainty in the parameters of interatomic potentials.

3. Formulate and validate surrogate models to evaluate the effects of uncertainty in interatomic potential parameters on materials properties predicted by MD simulations.

4. Utilize surrogate models to accelerate the solution of the inverse problem to determine the parameters of the interatomic potential based on observables.

1.4 Thesis Layout

Following the thesis objectives, this work is organized in the following way:

Chapter 2 presents the complete methodology followed in this thesis, it includes some background information for all the different techniques used within this work. First, the formulation of MD is described, followed by concepts and methods to propagate and quantify uncertainty in the interatomic potential parameters. Finally this Chapter provides an overview of the surrogate models used to describe the sensitivity of material properties due to the parametric uncertainty in the interatomic potential.

Chapter 3 gives a complete rundown of the methodology described in Chapter 2 for graphene material systems. Initially, the complete topology of the material, including the addition of defects in the system is described. A breakdown of the interatomic potential is given prior to detailing the mechanical properties and providing observations on their sensitivity. In general this chapter aims to provide insights on how this material behaves when simulated using the Adaptive Intermolecular Reactive Empirical Bond Order (AIREBO) potential.
Chapter 4 follows a similar layout as its predecessor chapter. This chapter deals with a metallic material system, specifically aluminum, and provides a different application of the methodology using the more common Embedded Atom Model (EAM) interatomic potential.

Chapter 5 presents a feasibility study to accelerate a class of computationally expensive inverse problems in MD by using surrogate models. The inverse problem aims to estimate the parameters of the AIREBO interatomic potential based on a given set of pseudo-experimental observables. Additionally, this chapter presents a study to solve the inverse problem when experimental noise is introduced to the observables.

Chapter 6 presents the final remarks by drawing conclusions based on the aforementioned chapters. In addition, it provides recommendations for future research work.
Chapter 2

Methodology

This chapter provides a complete description of the methodology followed in this thesis. Background is included for the general topics as well as for the mathematical aspects of uncertainty quantification. A complete definition of all terms and formulations is also included and will be used as background information for subsequent chapters.

2.1 Molecular Dynamics Simulations

Part of the main core of this thesis relies on computational modelling of material systems using MD, therefore this section covers the basic applications and theory involved in defining MD simulations.

2.1.1 Background and Applications

As materials and biological systems become more complex, new and more sophisticated methods are necessary to understand their properties and behaviour. At the same time, computers have become increasingly more powerful, providing a perfect resource for computationally modelling these systems, shedding some light on their characteristics as well as providing insights on how they work or interact. In the late 1950s a group of researchers investigating the dynamics of simple liquids developed a technique for modelling atomic behaviour based on classical mechanics [22]. Simulations that followed in the early 1960s combined improved numerical methods for integration with more realistic potential functions to study the dynamics of radiation damage or motion of liquids [23].

Currently, many techniques and algorithms can be implemented to consider quantum, temperature, and pressure effects to simulate systems in a more realistic manner. However, the more detailed the simulation, the more computational resources it requires,
making it difficult to find an optimum balance between performance and accuracy. It is still obvious that MD simulations offer a method to understand the dynamical properties of a system such as transport coefficients, time-dependent responses, and rheological properties. In fact, classical MD simulations are currently employed in the study of protein synthesis, liquid solutions [24], and a diversity of complex material systems. These techniques are widely used in combination with experimental procedures such as nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography. In the field of physics, MD allows researchers to understand the dynamics of systems at an atomic-level that would otherwise be impossible to observe directly, such as thin film growth and ion-subplantation. Similarly, MD aids in the research of new advances in nanotechnology by simulating structures that have not been created to understand their stability and behaviour. MD is also useful in the evaluation of new technologies such as electrical devices, nano-indentation, material removal mechanism, and process parameters such as cutting speed and forces [25].

2.1.2 General Theory

Classical MD is a computational method to describe the interaction between atoms or molecules. Each atom is modelled as a point mass where Newtonian physics is used to integrate the equations of motion, and solve the many-body problem by updating velocities, positions, and forces of the system at each time step. The MD model requires information from its constituents (i.e., atoms, molecules, surfaces, etc.) and must obey theoretical constraints such as energy conservation laws and boundary conditions [17].

As described by [26], carrying out MD simulations is very similar to performing real experiments. During physical experimentation, there is a sample piece connected to a measurement instrument, and a property of interest is measured over a period of time. In MD simulations, the sample consists of the topology of the material system and techniques to solve the required equations until the system reaches an equilibrium state over some designated time in order to calculate the desired properties. Similar sources of error can arise in both simulations and experiments. For instance, the sample is not prepared correctly, measurements are not carried out at the right settings or for the proper amount of time, or the system undergoes irreversible changes during the execution of the measurement.

Performing MD simulations requires three important ingredients; firstly, the model that describes the interaction between constituents must allow, given the atomic positions, the calculation of the associated potential energy, the atomic forces, and any
required field along the boundaries [27]. The relationship between the constituents is generally unknown but can be approximated by force fields or interatomic potentials. Secondly, an algorithm for numerical integration is needed to propagate atomic positions and velocities from time $t$ to $t + \delta t$. This step requires careful consideration when choosing the time step in order to guarantee stability of the integrator [28]. Finally, the use of a statistical ensemble is required to control pressure, temperature, and the number of particles in the system.

**Equations of Motion**

Since the force field or interatomic potential is what describes the interaction between atoms, it is possible to calculate the force acting on the $i$th atom as:

$$F_i = -\nabla_{r_i} U(r_1, \cdots, r_N) \quad (2.1)$$

where $U$ represents the potential energy of $N$ interacting atoms as a function of atomic positions $r_i = (x_i, y_i, z_i)$. Once the atomic forces are known it is possible to calculate the acceleration of each atom using Newton’s second law as follows:

$$m_i \ddot{r}_i = F_i \quad (2.2)$$

Since it is important to keep track of velocities as well as positions, the equations of motion can be written as a system of coupled differential equations that include the many-body interactions [29]:

$$m_i \frac{dv_i}{dt} = \sum_j F_2(\vec{r}_i, \vec{r}_j) + \sum_j \sum_k F_3(\vec{r}_i, \vec{r}_j, \vec{r}_k) + \cdots \quad (2.3)$$

where $F_2$ and $F_3$ represent the two and three body interactions respectively and many-body interactions can be added if necessary. In terms of computational resources, MD simulations are not memory intensive, since only vectors of atomic information are stored. Despite their efficiency, there are still limitations to MD simulations, as described later in section 2.1.3. Furthermore, some truncation is allowed between atomic interaction when atoms are too far apart (cutoff distance) and the remaining forcing terms are constructed to include some quantum effects [15]. It is important to point out that the system of equations 2.3 are stiff, meaning that any algorithm implemented to integrate the system must cope with short and long timescales.

It is necessary to march these equations over time by defining a time step ($\Delta t$) and
updating the required information over this time step. One of the most common methods to carry out time integration is the Verlet algorithm [30]. Although there are many versions of this algorithm, the most basic implementation is the velocity Verlet algorithm as follows:

\[
\begin{align*}
    v(t_0 + \Delta t) &= v(t_0) + \frac{dv}{dt}\bigg|_{t_0} \frac{\Delta t}{2} \\
    r(t_0 + \Delta t) &= r(t_0) + v(t_0 + \frac{\Delta t}{2}) \Delta t \\
    v(t_0 + \Delta t) &= v(t_0 + \frac{\Delta t}{2}) + \frac{dv}{dt}\bigg|_{t_0} \Delta t
\end{align*}
\] (2.4)

By applying these steps repeatedly, it is possible to map the atomic trajectories based on specified initial conditions. Given a small time step, this method conserves the energy reasonably accurately, with errors in the order of $10^{-7}$. Additionally, it can be proven that the method is of the order $\mathcal{O}(\Delta t^2)$.

**Temperature and Pressure Controls**

When performing MD simulations some degree of control is necessary to keep track of the evolution of atomic properties. Similar to physical experimentation, it is desirable to adjust the state of the simulation. For instance, controlling the density is achieved by setting the volume of the sample in the simulation. In practice, controlling parameters such as temperature and pressure is most advantageous [26].

Temperature control is achieved by adjusting the kinetic energy of the system, in particular by scaling the atomic velocities using various thermostat algorithms such as Anderson, Berendson, and Nose-Hoover [31]. Thermostats make use of the following relationship in an N-body system:

\[
T = \frac{1}{3Nk_B} \sum_i m_i v_i^2
\] (2.5)

where $k_B$ is the Boltzmann’s constant and $T$ is the instantaneous temperature of the system. Due to velocity fluctuations, it is not possible to keep this measurement of temperature constant throughout the simulation. Thus, thermostats work by maintaining the average temperature of the system constant.

Pressure control is achieved based on the generalized equipartition theorem and is defined, for the simplest interaction, as follows [32]:

\[
P = \frac{1}{V} \left[ Nk_B T + \frac{1}{3} \left< \sum_i r_i \cdot F_i \right> \right]
\] (2.6)

Similar to the temperature control, average values are kept constant by barometric algo-
rithms such as the Berendsen or the Nose-Hoover barostat. Barometric conditioning is relatively more complicated than temperature control, as it requires adjustment of the volume in the simulation. This adjustment demands performing several short simulations to keep pressure changes smooth during the simulation process.

Depending on the property to be calculated, MD simulations make use of ensembles to collect thermodynamic states. The most common ensembles used in MD are: the microcanonical, the canonical, and the isobaric-isothermal ensembles [33]. The microcanonical ensemble is known as the constant-energy, constant-volume ensemble or NVE ensemble, and is obtained by conserving energy when solving the equations of motion without any pressure or temperature control. NVE ensembles are not very common when reaching system equilibrium, as they do not allow control of any kind. Canonical ensemble consists of a constant-temperature, constant-volume or NVT control. The NVT ensemble makes use of thermostat algorithms to control temperature while maintaining a constant volume throughout the simulation. The final ensemble is the isobaric-isothermal or NPT ensemble, which consists of using a thermostat together with a barostat to maintain pressure and temperature at a constant stage. Each of the aforementioned ensembles have their advantages and disadvantages and careful considerations must be applied when using them during MD simulations.

**Boundary Conditions**

The advantage of performing MD simulations is the ability to obtain bulk material properties without the need of highly expensive experimental methods. To do so, periodic boundary conditions (PBC) are used to simulate an infinitely large system. For any system, constructing PBC implies replication of the simulation box in all directions in such a way that if a particle leaves the environment in one direction, another particle will enter from the opposite direction. Figure 2.1 shows the implementation of the PBC applied to a 2D system [15]. The central region depicts the simulation box and the arrows represent the direction of motion for each particle. By using PBC no boundary effects are present in the system allowing the reproduction of bulk properties.

**2.1.3 Limitations**

Although MD simulations offer a reliable method to study the properties of material systems, there are still many limitations to this method. The first limitation introduced in MD is the use of classical forces to define atomic movement. Even though MD uses Newtonian physics, systems at the atomic level obey the laws of quantum mechanics and
follow the Schrodinger equations [32]. However, this classical approximation becomes valid as long as the de-Broglie thermal wavelength is much smaller than the mean nearest neighbour separation. Current state of the art MD simulations can be developed to work simultaneously with \textit{ab initio} calculations to incorporate some quantum effects.

Two more significant issues are the time and size scales. MD simulation are usually limited to time scales ranging anywhere from $10^{-12}$ up to $10^{-9}$ seconds which does not cover a whole range of chemical or physical processes such as slow diffusion, protein folding, or the dynamics of polymer systems [28]. Similarly, limitations with length scales are also present as it is computationally challenging to model systems beyond a few tens of nanometers. For many materials, this presents a serious practical limitation since many macro-scale properties such as hardness, surface tension, and thermal diffusion and expansion do not correlate directly to the atomic scales.

Yet another limitation, already mentioned in Chapter 1 and one that is part of the focus of this thesis, is the reliability of the interatomic potentials, i.e., how realistic are the atomic interactions of the system. The idea behind the interatomic potential is to mimic the forces experienced by real atoms; however, many of these interactions suffer from a truncation that modifies the complete interaction of the system. Interatomic potentials are created by applying curve fitting methods to a limited set of experimental data or results from density functional theory. However, the current fitting techniques do not account for uncertainty introduced during experimentation and can only describe accurately those material properties that were fitted with a high weight during the procedure. By shifting the weights to different material properties, a new interatomic potential can
be created; however, a single interatomic potential should be transferable and be able to capture all the properties of the material systems accurately.

2.2 Uncertainty Propagation and Quantification

It has become essential to understand the uncertainty that gives rise to errors when performing simulations to model real dynamic systems; as such, it is important to quantify and determine sources of uncertainty as well as to develop techniques to study their propagation. In a general sense, it is necessary to investigate how the variability in the input parameters of the model affect the output parameters or observables. This section describes the techniques used for the propagation and quantification of uncertainty throughout this thesis.

2.2.1 Background: Prediction of Errors

When working with MD simulations it is important to question: what are the sources of error that can limit the measurement of required observables? how can these errors be quantified and mitigated in a systematic way to improve the simulation results? To answer these questions, it is necessary to ensure that the “unknown unknowns” can be calibrated into the simulation model. In general, it is very difficult to quantify these unknowns; however, it is possible to quantify the effects a slight change in the input will have when applied to an advanced mathematical model.

According to Smith [34], uncertainty quantification (UQ) can be define as “the science of identifying, quantifying, and reducing uncertainties associated with models, numerical algorithms, experiments, and predicted outcomes or quantities of interest”. The field of uncertainty quantification uses diverse analysis methods and statistical tools to critically assess mathematical models and simulations. Nonetheless, it becomes critical to recognize the sources of uncertainty that prevail in computer simulations [34, 35]:

Parameter uncertainty: This relates to the input parameters that are difficult to estimate accurately or whose values cannot be inferred using statistical methods.

Model inadequacy: This source of uncertainty arises due to imperfections of the computer model itself. Even with a high degree of fidelity in the model, it is difficult to capture all the dynamics of the system precisely; therefore, the observables predicted by the model will not equal the true value of the process.
Parametric variability: Refers to inputs to the model that are uncontrolled, unspecified or require more detail than available. Usually, these parameters vary according to a statistical distribution, which introduces additional uncertainty to the system.

Algorithm uncertainty: Relates any machine uncertainty that is introduced in the system due to the algorithm implementation, this can include issues with machine precision as well as required configurations to the algorithm in order to obtain the desired output.

Although there are well developed fields to describe numerical errors in experiments or simulations, the objective of UQ is to systematically identify and measure errors in models or simulations to analyze how they propagate and affect predicted outcomes. Current UQ methodologies may involve an initial study of the input parameters combined with a sensitivity analysis to achieve a good parameter selection. This step is usually followed by techniques to reduce the biases or discrepancies in the model to obtain the proper calibration. Finally, uncertainty propagation from input parameters to the quantities of interest can be investigated [34]. Uncertainty propagation techniques can be classified as forward or inverse; where forward methods focus on the influence on the outputs from the parametric variability in the input parameters, while inverse methods aim to describe the distribution of the input parameters based on the a collection of observables obtained from simulations or experiments.

Uncertainty quantification helps identify the confidence intervals of the predicted observables; it is also a fundamental component for model validation, as it allows for a more robust design. Techniques to propagate uncertainty are based on either sampling methods, spectral representation or perturbation methods. However, due to the nature of MD applications, this thesis focuses on sampling methods, which are discussed in the following section.

2.2.2 Monte Carlo

Created by mathematician Stanislaw Ulam in the 1940s while working on the Manhattan project, the Monte Carlo method is the most common sample technique to propagate uncertainty. Earlier techniques to approximate errors were based on Taylor expansion under the assumption that the system to be analyzed could be linearized. In contrast, Monte Carlo sampling is a method in which linearization approximation is not required [36].

Monte Carlo techniques are a class of probabilistic, non-deterministic numerical methods that rely on random sampling and are used to compute statistical quantities when
analytical methods are not convenient [37, 38]. Applying Monte Carlo simulations to uncertainty analysis can help solve one or both of the following problems:

**Generating realizations:** Monte Carlo can help generate a single realization of \( \mathbf{x} \) from a probability distribution function \( p(\mathbf{x}) \).

**Statistical moments:** This technique can help evaluate statistical moments represented by integrals of the following form:

\[
I = \langle \phi(\mathbf{x}) \rangle = \int \phi(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}
\]

(2.7)

where \( \mathbf{x} \) is a vector of random variables with a known probability distribution function \( p(\mathbf{x}) \), and \( \phi(\mathbf{x}) \) is the function of interest [39].

The basic algorithm for Monte Carlo simulations is very simple in nature and requires: (1) drawing \( N \) samples of the random variables from their probability distribution functions, (2) evaluation of the function of interest at these points, and (3) estimating the integral in the following way:

\[
I \approx \bar{I} = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)})
\]

(2.8)

Since the vector of random variables \( \mathbf{x} \) is obtained from \( p(\mathbf{x}) \), the expectation of \( \bar{I} \) is \( I \). The accuracy of the Monte Carlo estimate is independent of the dimensionality of the space being sampled. The standard error for this method is given by \( \frac{\sigma_{\phi}}{\sqrt{N}} \), where \( \sigma_{\phi} \) is the standard deviation of the function of interest [40, 41] and can be obtained as follows:

\[
\sigma_{\phi}^2 = \int (\phi(\mathbf{x}) - I)^2 p(\mathbf{x}) d\mathbf{x}
\]

(2.9)

Monte Carlo sampling methods have the advantage of being easy to implement. However, these methods are not without limitations; for instance, these techniques require a large number of samples or realizations to be accurate. In fact, convergence in Monte Carlo methods is guaranteed due to the strong law of large numbers which is described as follows [42]: For a sequence of independent identically distributed (i.i.d.) random variables \( X_1, X_2, \cdots, X_n \) with expected value \( \langle X_i \rangle = \mu \) and \( S_n = X_1 + X_2 + \cdots + X_n \) the sequence of sample means \( \frac{S_n}{n} \rightarrow \mu \) as \( n \rightarrow \infty \).

An additional limitation of these methods is their dependency on a reasonable representation of the probability distribution function. For applications where the distribution
Chapter 2. Methodology

is assumed to be normal or uniform, these methods are highly efficient. However, for correlated, non-Gaussian, non-Uniform random variables the task becomes more laborious.

2.2.3 Sparse Quadrature

One of the disadvantages of Monte Carlo techniques is the need for a large number of realizations to reach convergence. When simulations become computationally expensive, techniques such as sparse quadratures can be used instead. The integrals that arise in uncertainty analysis can also be evaluated using quadrature methods. The most common examples are the Trapezoid or Simpson’s rule for 1D functions. In a more general sense, quadrature methods seek to approximate integrals of the form:

\[ I(f) = \int_a^b f(x)dx \]  

using \( n_l \) function evaluations, as follows:

\[ I(f) \approx I_l^{(1)}(f) = \sum_{i=1}^{n_l} c^{(i)} f(x^{(i)}) \]  

where \( I_l^{(1)} \) denotes an one-dimensional quadrature rule with level \( l \). \( c^{(i)} \) and \( x^{(i)} \) represent the quadrature nodes and the weights respectively \([1, 43]\). The level \( l \) is an indicator of the increase of number of nodes and accuracy for a given dimension. The generalization of the 1D quadrature rule allows the construction of a N-dimensional rule by using a combination of tensor products of this quadrature rule, i.e., a sum over all possible combinations of the nodes and weights. The N-dimensional rule is then given by:

\[ I^{(N)} f = (I_{l_1}^{(1)} \otimes \cdots \otimes I_{l_N}^{(1)}) f = \sum_{i_1=1}^{n_{l_1}} \cdots \sum_{i_N=1}^{n_{l_N}} c^{i_1}_{l_1} \cdots c^{i_N}_{l_N} f(x_1^{(i_1)}, \ldots, x_N^{(i_N)}) \]  

Full tensor quadrature techniques, such as the one described above, are simple to implement but offer a significant disadvantage, the “curse of dimensionality”. This simply means that the number of function evaluations grows exponentially as the dimension of the problem increases. Full tensor products involve \( \prod_{i=1}^N n_{l_i} \) function evaluations and are inefficient due to the large number of evaluations required. A more efficient approach for evaluating high-dimensional integrals is to use sparse tensor product quadrature rules, also known as sparse grids \([43]\). Using sparse grids results in the same accuracy but with fewer nodes, hence fewer function evaluations. In order to construct the sparse
grid, consider the one-dimensional quadrature rule shown in equation 2.11, and define the difference formulas:

\[
\Delta^{(1)}_l = (I^{(1)}_l - I^{(1)}_{l-1}); \quad I^{(1)}_0 = 0
\]  

(2.13)

Introducing a multi-index \( l = (l_1, l_2, ..., l_N) \in \mathbb{N}^N \) which indicates the level in the multivariate sense and defining \( |l| = \sum_i^N l_i \) then the sparse cubature at level \( l \) is defined as:

\[
I^{(N)}_l = \sum_{|l| \leq l+N-1} (\Delta^{(1)}_{l_1} \otimes \cdots \otimes \Delta^{(1)}_{l_N})
\]

(2.14)

The sparse grid method can exploit the smoothness of the function \( f \) and produces an integration error of the following order:

\[
O(M^{-r}(\log M)^{(r+1)(N-1)})
\]

(2.15)

where \( M \) represents the total number of quadrature points and \( r \) is the smoothness of the integrand. Obtaining the total number of points in the sparse grid can be achieved by counting the additional points included when moving from level zero to the next level until reaching the required level \( L \). Since the order for the 1D case is given by:

\[
n^1_l = 2^l + 1; \quad \text{for } l = 1, \cdots, L; \quad \text{and } n^1_1 = 1
\]

(2.16)

the additional points can be calculated, for the 1D case, as the difference between the order of the current and the previous levels, and can be referred as the sequence of \( new_{1D} \) points. For levels \( l = 2 \) onwards, this can be obtained by \( 2^{l-1} \). Using the level vector \( l \) the total number of points is given by:

\[
n^N_l = \sum_{l=0}^L \sum_{|l|=l} ^N \prod_{i=1}^N new_{1D}(l(i))
\]

(2.17)

which is a sum over all product rule levels from 0 to \( L \), of the sum over all product grids of level exactly \( l \), of the number of additional points in that product grid. In equation 2.17 the index \( i \) refers to the element of vector \( l \), which in turn refers to the element in the sequence \( new_{1D} \). For instance, the number of points of the sparse grid of level 3 and dimension 2 is given by:

\[
n_3^2 = (1 \cdot 1) + (2 \cdot 1 + 1 \cdot 2) + (2 \cdot 1 + 2 \cdot 2 + 1 \cdot 2) + (4 \cdot 1 + 2 \cdot 2 + 2 \cdot 2 + 1 \cdot 4)
\]

\[
= 29
\]

(2.18)
A comparison between the full tensor product and sparse grid is given in Figure 2.2. Both quadratures are able to integrate a polynomial of order 16 exactly. However, the full quadrature requires 224 more points to achieve this task, almost a 440% increase from the sparse tensor product. This advantage makes sparse quadratures more desired for applications with large number of dimensions.

![Figure 2.2: Comparison between full tensor product and sparse tensor product to integrate a function of order 16 exactly. Full quadrature requires 289 points while the sparse quadrature requires only 65 [1].](image)

### 2.2.4 Latin Hypercube Sampling

Another sampling method that is efficient when large dimensionality is present is the Latin Hypercube Sampling (LHS). This method is a form of stratified sampling that becomes effective when one dimension is particularly important. As pointed out by Huntington [44], LHS preserves probability distributions for each of the random variables, while matching their target correlations. In order to do this and to allow for a small number of samples, LHS constructs a highly dependent joint probability density function for the random variables in the problem.

Latin Hypercube sampling is an simple technique to implement and works well in combination with surrogate models to examine the sample space. For the case of uniform distributions, LHS generates sample points as follows [45]:

\[
X_{ij} = \pi_j(i - 1) + U_{ij}, \quad 1 \leq i \leq n, \ 1 \leq j \leq d
\]  

(2.19)

where \(X_{ij}\) is the random realization from \(n\) number of samples and \(d\) dimensions. Simi-
larly, $\pi_j$ are uniform random permutations of $\{0, 1, \cdots, n - 1\}$, $U_{ij} \sim U[0, 1)$, and all the $U_{ij}$ and $\pi_j$ are independent. Figure 2.3 shows a comparison between 10 points obtained from a random sample versus the stratified Latin Hypercube sampling technique in two dimensions $X_1$ and $X_2$.

![Comparison between a random sampling method and Latin Hypercube sampling](image)

Figure 2.3: Comparison between a random sampling method and Latin Hypercube sampling. Using $n = 10$ points for two dimensions $d = 2$ ranging from $[0,1]$

### 2.3 Surrogate Model

As mentioned in Chapter 1, one of the objectives of this work is to formulate models where conclusions on the sensitivity of the material properties can be drawn. Surrogate or metamodels aim to build approximation models to high fidelity systems to reduce programming complexity, computational cost, and operational time. This sections describes the surrogate models being implemented as well as a description of the model validation and its use for sensitivity analysis.

#### 2.3.1 Background and Applications

The use of computational models in many science and engineering applications can provide reliable methods for design iterations that would otherwise require expensive and time consuming experimentation. In cases where large number of iterations are required, high fidelity computational models can also hinder or delay the design process. In such cases, the construction of simple approximation models capable of predicting system performance could be the optimal solution. Surrogate modelling is a method to construct such approximations, where a relationship between input and output data is developed
to accurately mimic the behaviour of the high fidelity model while being computationally cheaper to evaluate [46].

As described in [47], if a high fidelity model is represented by a functional form of the type $y = f(x)$, then objective of a surrogate model is to find the approximation $\hat{y} = \hat{f}(x, \alpha) \approx f(x)$. In this case, $x \in \mathbb{R}^p$ represents a vector of input values, $y$ is the observable value, and $\alpha$ is a vector of unknown parameters which is estimated using either a black-box approach or an intrusive approach. This work focuses on black-box surrogate models, where a few steps must be followed in order to construct the surrogate.

1. **Design of Experiments (DoE):** This is the process of collecting or generating data to train and test the model. In general, the data points can be selected using any of the techniques described in the previous section. For a black-box approach, it is usually necessary to generate the input parameters and run them through the black-box to obtain the output data, thus obtaining an input-output dataset.

2. **Model Structure and Parameter Estimation:** This step ensures the selection of the correct model type in order to estimate the unknown parameters accurately. For instance, linear, quadratic or general non-linear models can be selected to represent the approximation function and minimization strategies can be used to reduce errors in the parameter estimation. Some of the most popular models include Kriging or Gaussian process models (GPMs), Radial Basis Functions (RBFs), and support vector machines.

3. **Model Assessment:** In this step, the accuracy of the surrogate model is evaluated. This process can be performed by applying classical techniques, such as goodness-of-fit and diagnostic plots.

The iteration of the steps mentioned above is crucial for the creation of the surrogate model. Further implementation of the model can serve to search the design space or optimize the design process.

### 2.3.2 Bayesian Gaussian Process Modelling

Applying Bayesian statistics is very common in the creation of surrogate models because it allows for the adaptation and update of the model based on information obtained from it. In essence, the Bayesian model allows for the combination of new data obtained from the model with the existing knowledge of the model. The Bayesian approach is based on Bayes’ rule where a posterior probability can be inferred based on a prior probability
and a likelihood probability; in general the rule can be written as [48]:

\[
P(H|E) = \frac{P(E|H)P(H)}{P(E)}
\] (2.20)

where \(P(H|E)\) is the posterior probability or the probability of the hypothesis \(H\) given the evidence \(E\). \(P(E|H)\) is the probability of observing \(E\) given \(H\), this is also called the likelihood. Finally, the terms \(P(H)\) and \(P(E)\) are the prior probability (probability of the hypothesis occurring before the evidence is obtained) and the model evidence respectively.

The black-box stochastic model presented in this section is based on the Bayesian Gaussian process modeling. It is introduced in order to identify the sensitivity of the observables in terms of the input parameters described by the interatomic potentials being studied. The objective of creating this model is to obtain a computationally efficient method for detailed parametric investigations [49]. In terms of the Bayesian formulation, the Gaussian process prior is used to infer a posterior process that is also Gaussian. The stochastic model structure is represented as [47]:

\[
Y(x) = \beta + Z(x)
\] (2.21)

where \(Y(x)\) is a vector containing the observables as a function of the vector of input parameters \(x\). \(\beta\) is an unknown hyperparameter that can be obtained from the data, and \(Z(x)\) is a Gaussian stochastic model process with zero mean and covariance given by:

\[
\text{Cov}(Z(x, x')) = \Gamma(x, x') = \sigma_z^2 R(x, x')
\] (2.22)

here, \(R(x, x')\) is a parametrized correlation function that can be tuned from the dataset and \(\sigma_z^2\) is the process variance, which can also be estimated from the data. In order to use this statistical framework, it is assumed that the observed outputs are realizations of a Gaussian random field with mean \(\beta\) and covariance function \(\Gamma\). As stated in [47], a common correlation function used is given by:

\[
R(x, x') = \prod_{j=1}^{p} \exp(-\theta_j |x_j - x'_j|^{m_j})
\] (2.23)

where \(\theta_j \geq 0\) and \(0 < m_j \leq 2\) are undetermined hyperparameters. The value of \(m_j = 2\) can be thought as a smoothness parameter and is kept at this value following the assumption that the function being modeled is infinitely differentiable. The value of \(\theta_j\) in
Chapter 2. Methodology

22

this definition accounts for the non-linearity of the model. Using the Bayesian approach, two levels of inference can be made to determine the unknown hyperparameters, and to obtain a prediction value of the observables at a new point. By applying Bayes’ rule as the first level of inference, it is possible to determine the hyperparameters given the dataset as follows:

\[
P(\theta, \beta, \sigma^2_z|D) = \frac{P(D|\theta, \beta, \sigma^2_z)P(\theta, \beta, \sigma^2_z)}{P(D)}
\] (2.24)

where \(D\) is the set of input-output parameters. Once the hyperparameters are determined, the second level of inference uses the aforementioned fact that the posterior distribution of the observables, that is, \(P(y(x)|D, \theta, \beta, \sigma^2_z)\) is also Gaussian. In general, the prediction at an unseen point \(x\) can be written as \(Y(x) \sim N(\hat{Y}(x), C(x, x'))\) where the posterior mean and covariance are given by:

\[
\begin{align*}
\hat{Y}(x) &= \beta + r(x)^T R^{-1}(y - 1\beta) \\
C(x, x') &= \sigma^2_z(R(x, x') - r(x)^T R r(x'))
\end{align*}
\] (2.25)

where \(R\) is the correlation matrix whose \(ij^{th}\) element is given by \(R(x_i, x_j)\). The correlation between \(x\) and all points in the dataset is captured by \(r(x)\), such that \(r(x) = \{R(x, x_1), R(x, x_2), \ldots, R(x, x_n)\}\). The vector \(y\) contains all the observables in the training set, and the vector \(1\) is the unity vector of the same size as the observables vector. As noted by [47], for computational efficiency, it is possible to compute the Cholesky decomposition of \(R\) yielding only a vector-vector product so that, \(\hat{Y}(x) = \beta + r(x)^T w\); where \(w = R^{-1}(y - 1\beta)\).

2.3.3 Radial Basis Function Surrogates

During the introduction of this section, the goal of the surrogate model was defined as the approximation the function of interest, i.e., \(\hat{y} = \hat{f}(x, \alpha) \approx f(x)\). However, there has been no definition on the form of \(\hat{f}\). The simplest form of this approximation is made by a generalized linear model of the form [37, 47]:

\[
\hat{f}(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)
\] (2.26)
where the $φ_i$ represent the set of basis functions for $i = 1, 2, \cdots, N$. Radial basis functions (RBF) use a particular set of basis functions of the form:

$$φ_i = K(∥x - µ_i∥)$$  \hspace{1cm} (2.27)

The function $K$ is known as the kernel, the value $µ_i$ represents the centre of the basis function, and the $∥·∥$ represents the Euclidean norm. Depending on the application, the RBF can take several forms such as linear or cubic splines, Gaussian, or multiquadratics. In this work, a Gaussian kernel function of the form shown in equation 2.28 will be used.

$$K(∥x - µ_i∥) = \exp \left( -\frac{∥x - µ_i∥^2}{θ} \right)$$  \hspace{1cm} (2.28)

where the shape parameter $θ$ can be used to control the domain of influence of the RBF [47]. From equations 2.26 and 2.28 it can be noted that the next step in developing this model is to compute the weights $α_i$ and estimate the shape parameter. To compute the weight vector it is necessary to ensure that the model matches the observables given from the training dataset. This allows for the formulation of a linear system of equations given as follows:

$$y = Kα$$  \hspace{1cm} (2.29)

where $K$ is the Gram matrix formed using the training dataset where the $ij^{th}$ element is given by $K_{ij} = K(∥x_i - x_j∥)$. The use of positive-definite kernels guarantees that the Gram matrix is non-singular, such as in the case of Gaussian kernels [47]. However, since the Gram matrix is dense, capturing the weight vector requires $O(N^2)$ and $O(N^3)$ for memory and operations respectively.

Estimation of the shape parameter is crucial to obtain a good prediction for the testing dataset. A good method to achieve this is to minimize a predictor function that acts as the interpolation error. For instance, it is possible to use a maximum likelihood estimator (MLE) to estimate the shape parameter; in this case maximizing the likelihood function is similar to minimizing the following expression:

$$y^T K^{-1} y \cdot \left[ \det(K) \right]^{1/N}$$  \hspace{1cm} (2.30)

Note that as $N$ increases the value of this expression approaches 0, making numerical errors more likely to arise. To avoid such errors, an identity involving the eigenvalues $\lambda$
of the Gram matrix can be invoked as [50]:

\[
det(K) = \prod_{i=1}^{N} \lambda_i(K)
\]  

(2.31)

Computations for calculating the eigenvalues for the Gram matrix are of order \(O(N^3)\). Similarly, computing the inverse of the \(K\) matrix is of order \(O(N^3)\) and it should be noted that the inverse matrix can be obtained from equation 2.29 since \(K^{-1}y = \alpha\).

A different method of estimating the shape parameter of the kernel function is to use the leave-one-out (LOO) approach (See section 2.3.5), where the loss or prediction error using the sum of squares of errors is given by:

\[
LOO(\theta) = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{N} \alpha_j K(\|x_i - x_j\|) \right|^2
\]  

(2.32)

The main disadvantage with this methodology is that the order of computations reach \(O(N^4)\) when solving the \(N\) linear system of equations. However, Rippa [51] found that computational complexity can improve to \(O(N^3)\) by simplifying the error term:

\[
y_i - \sum_{j=1}^{N} \alpha_j K(\|x_i - x_j\|) = \frac{\alpha_i}{K_{ii}^{-1}}
\]  

(2.33)

where \(K_{ii}^{-1}\) is the \(i^{th}\) diagonal element of the inverted Gram matrix. By applying the LOO approach, it is possible to obtain a good estimation of the shape parameter and induce less errors than the MLE approach.

### 2.3.4 Reduced Order Modelling

The advantage of surrogate models is to provide a computationally fast method to describe complex systems that would otherwise take a long time to simulate. In many cases, these systems are represented by a large number of input or output parameters to which the creation of surrogates presents no real advantage. However, techniques such as reduced order modelling (ROM) can help decrease the dimensionality of the design space while retaining the underlying structure and dynamics of the data, ultimately lowering the computational complexity of the problem. From the different ROM techniques, this work employs the principal component analysis (PCA) to reduce the input space.

Principal component analysis can be defined as a linear combination of observed variables that are weighted in such a way that the resulting components account for the
maximum amount of variance in the dataset. In simple terms, PCA works by finding the orthogonal basis that maximizes the variance from the dataset and can be calculated by eigenvalue decomposition of the data correlation matrix [52, 53]. To formulate the problem, the input data matrix $X \in \mathbb{R}^{m \times n}$ with $m$ variables and $n$ realizations requires a change of basis such that:

$$Y = PX$$  \hspace{1cm} (2.34)

where $P \in \mathbb{R}^{m \times m}$ is the transformation matrix. For this process, it is also assumed that the input dataset is centred with respect to its mean, so that each sample set has a mean of zero. From the input dataset, it is possible to compute the covariance matrix as follows:

$$C_X = \frac{1}{n-1}XX^T$$  \hspace{1cm} (2.35)

where each of the entries in $C_x$ is the covariance pair between the $m$ variables. PCA requires that the variables in the transformed matrix are as uncorrelated as possible, in other words, the goal is to construct the covariance matrix of the transformed data $C_Y$ such that the variance is maximized while the covariance between variables is minimized. Using a similar formulation to equation 2.35, the covariance matrix of the transformed system is given by:

$$C_Y = \frac{1}{n-1}YY^T$$

$$= \frac{1}{n-1}(PX)(PX^T)$$

$$= \frac{1}{n-1}P(XX^T)(P^T)$$  \hspace{1cm} (2.36)

The matrix $XX^T$ is symmetric and therefore, it is possible to diagonalize it; i.e., the matrix can be written in the form $XX^T = EDE^T$ where the $D$ and $E$ are the diagonal matrix of eigenvalues and the corresponding eigenvector matrix respectively. To ensure the aforementioned relationship of maximizing the variance, the choice of the transformation matrix is taken such that $P = E^T$. From the results, it is evident that the principal components of $X$ are the rows of the transformation matrix $P$. Other techniques, such as singular value decomposition, can also be applied to compute the principal components without the construction of the covariance matrix; the reader is referred to [52] for further details.

### 2.3.5 Surrogate Model Validation

Model validation ensures that the surrogate model can make accurate predictions of the quantities of interest. Different techniques such as k-fold or leave-p-out (LpO) cross-validation, have been introduced to accomplish this task. The surrogate models used in
this work, can be validated using the leave-one-out cross-validation technique combined with a diagnostic to check the error estimate of the metamodel called the standardized cross-validation residual (SCVR) approach. To further estimate the validity of the metamodel, the leave-p-out approach, constructed from the basis of the leave-one-out method, is used to ensure the model is adequate for the problems in hand.

The leave-one-out is an exhaustive cross-validation method used to show a comparison between the actual observable values and the predicted values using the surrogate. It is constructed by leaving the $i^{th}$ training point out, testing the model at the removed point, and repeating for all the data points in the training set [47, 54, 55]. The predicted value is then compared with the actual value at the removed point to estimate the error.

In terms of the Gaussian process model, the leave-one-out procedure aims to predict the error in terms of the posterior variance ($\sigma^2 = C(x, x')$); however, the procedure can significantly underestimate the actual prediction error [56]. Therefore, it is used in combination with the SCVR approach to check the validity and accuracy of the model. The SCVR approach for the $i^{th}$ testing point is given by:

$$SCVR_i = \frac{y(x^{(i)}) - \hat{y}_{-i}(x^{(i)})}{\sigma^2_{-i}(x^{(i)})}, \quad i = 1, 2, \cdots, n$$

(2.37)

where $\hat{y}_{-i}(x)$ and $\sigma^2_{-i}(x)$ are the mean and variance of the model prediction at a point $x$ without using the $i^{th}$ training point. A good model prediction will have its SCVR points lie in the interval between $[-3, +3]$ which implies that the prediction lies between the $\pm 3\sigma$ from the mean value.

In a similar fashion to the LOO method, the LpO approach is constructed by using $p$ observables as the test set and the reminder of the data as a training set. The advantage of this exhaustive method is the fact that it is iterated over all possible combinations of $p$-observations with the corresponding training sets. However, this can also become a disadvantage because the construction of a new model is required for each of the observables and each of the possible combinations of training and testing sets.

### 2.3.6 Sensitivity Analysis

Based on the surrogate model created, a sensitivity analysis can be performed to gain insights into the input-output relationship of the data. The purpose of this analysis is to test the robustness of the model in the presence of uncertainty. Similarly, this analysis will determine which input parameters greatly influence the observables.

The idea is to isolate the effects of each of the input variables on each of the quantities
of interest. By integrating out the effects of other input variables, the main effect of a single variable \( x_i \) can be found by [47]:

\[
M(x_i) = \frac{1}{V} \int y(x) \prod_{j \neq i} dx_j
\]  

(2.38)

where \( V \) is the volume of the design space. Note that the integral uses \( y(x) \) to represent the observable value; however, the use of the approximation model \( \hat{y}(x) \) can be used to compute the main effects efficiently. For Gaussian process models, it is possible to analytically integrate the above equation as long as the correlation function \( R(x, x') \) obeys a product correlation rule [47]. When this is not the case, sampling techniques such as Monte Carlo or Latin Hypercube Sampling must be employed to approximate the integral. For the case of the Gaussian model presented in this thesis, with correlation function shown in the equation 2.23, these values are readily available due to the posterior calculations made by the surrogate model. Graphically, main effects tend to produce a flat response when the quantity of interest is insensitive to the input parameter.

In a similar fashion, the joint or interaction effects can be computed by further integrating out the contribution of all other input variables except the \( x_i \) and \( x_j \) inputs as follows:

\[
M(x_i, x_j) = \frac{1}{V^2} \int y(x) \prod_{k \neq i, j} dx_k
\]  

(2.39)

Global sensitivity indices are used as factors to further condense the main effect values [57]. The sensitivity indices are calculated in terms of their main effects as:

\[
S_i = \frac{\int M(x_i)^2 dx_i}{\sum_{j=1}^{n} \int M(x_j)^2 dx_j}
\]  

(2.40)

which represents the contribution of the variance of the main effect of \( x_i \) standardized with the total variance. Note that this implies that the sum of \( S_i \) values always add up to one.
Chapter 3

2D Material System: Graphene

This chapter applies the methodology described in the previous chapter to graphene material systems. First, the topology of the material is described followed by a description of the interatomic potential used in MD simulations. Finally, this chapter discusses the results of introducing uncertainty in the interatomic potential as well as sensitivity analysis of the quantities of interest.

3.1 Graphene Topology

Graphene is a two-dimensional material that was developed in 2004 by Kostya Novoselov and Andre Geim at the University of Manchester. This one-atom thick material was extracted from graphite by pulling out its layers and depositing them on SiO$_2$ using a process called the Scotch tape technique [58]. Due to this discovery, Novoselov and Geim were granted the Nobel Prize in Physics in 2010.

The topology of planar and pristine graphene is based on carbon atoms arranged in a two-dimensional benzene-ring structure. This densely packed sp$^2$-bonded carbon allotrope with a hexagonal lattices has a carbon-carbon distance of 0.142 nm and exhibits rather promising properties that could be used in electronics or structural applications.

Graphene is considered the building block of graphitic materials; in other words, it is the basic structural element of other allotropes such as graphite (stacking graphene layers together), carbon nanotubes (rolling graphene sheets), charcoal, and fullerenes [59]. As such, graphene research has become increasingly valuable as scientists aim to determine its exact properties, composition, and structure.
3.1.1 Defects in Graphene Topology

Using different visualization techniques such as scanning tunnelling microscopy (STM) or transmission electron microscopy (TEM), research has shown that graphene exhibits many types of defects [60]. These defects can alter the mechanical or electrical properties of the material as they change the physical and chemical properties of the carbon structure.

In order to understand defects in graphene, a complete survey including their formation and influence in material properties has been presented by Banhart [61]. The survey is motivated by the fact that the extraordinary properties of graphene have been confirmed but only for structures with low density of defects. Stone-Wales, single and double vacancies, and line defects are depicted to show how the graphene structure accommodates these defects by changing from hexagon rings to other energetically favourable configurations such as heptagon-pentagon rings in order to become stable. Other defects pointed out include the addition of atoms which are shown to use the third dimension, meaning that the new atom will create covalent bonds with other carbon atoms but will be situated on a different plane as the whole graphene sheet as this becomes energy favourable.

Efforts to determine the mechanical properties of graphene under defects have been successful. For instance, using atomistic-based continuum progressive fracture models, based on finite element analysis, it was possible to investigate the tensile strength of graphene containing vacancies [62]. The study shows how detrimental vacancies can be for the mechanical properties of graphene. In fact, a low density of 4.4% of missing atoms (10 vacancies in a 217 atoms graphene sheet) shows a reduction in the tensile strength of graphene of 50%. Furthermore, dislocations also have an effect on the electrical and magnetic properties of graphene. Dislocations that form heptagon-pentagon rings greatly affect the magnetic properties of the system as they introduce distortions in the local density of states [63].

Using high-resolution microscopy it has been possible to determine the polycrystalline nature of graphene. The inclusion of grain boundaries in graphene sheets has fuelled significant research to determine how grain size and configuration affects the properties of the material. Similar to vacancies, having a network of grain boundaries reduces the strength of graphene by about 50% [64]. Additionally, it was found that larger grain sizes ($\gg 15$ nm) could weaken the material in a larger scale as cracks start to propagate through the bulk of the neighbouring grains [65]. The significance of these defects in defining the material properties is the reason why they are explored in this thesis.
3.2 AIREBO Interatomic Potential

Prior to describing the MD simulations used to define the material properties of graphene, this section will focus on introducing the Adaptive Intermolecular Reactive Empirical Bond Order (AIREBO) interatomic potential used for such a task. This potential is created by the combination of three potentials: the original REBO potential which accounts for repulsive and attractive forces between atoms in the system, a LJ term to account for the intermolecular interaction that include the dispersion and short-range repulsion effects, and a TORSION term dependent on dihedral angles to account for torsional interactions between single bonds [66].

The AIREBO potential was mainly developed to complement the predecessor REBO potential. The addition of non-bonded interactions make it suitable for systems with high intermolecular interactions such as graphite, as well as covalent systems such as graphene. The potential also includes a four-body torsional interaction which becomes important in many curved carbon systems. The general form of the AIREBO potential is given by:

\[
E = \frac{1}{2} \sum_i \sum_{j \neq i} \left[ E_{ij}^{\text{REBO}} + E_{ij}^{\text{LJ}} + \sum_{k \neq i} \sum_{l \neq i,j,k} E_{ijkl}^{\text{TORSION}} \right]
\]  

(3.1)

where \(E\) is the total potential energy of a system with atoms \(i, j, k,\) and \(l\). The REBO potential which describes the intra-atomic interactions in carbon is described as:

\[
E_{ij}^{\text{REBO}} = V_R^{ij}(r_{ij}) + b_{ij}V_A^{ij}(r_{ij})
\]

(3.2)

where \(V^R\) and \(V^A\) are the repulsive and attractive part of the potential respectively; \(b_{ij}\) is the bond order which allows for changes in the strength of the bond depending in the local environment. For instance, this number is large for carbon-carbon interactions that follow a \(\text{sp}^2\) bond such as graphene. In this case, \(r_{ij}\) is the distance between atoms \(i\) and \(j\) up to a cut-off distance.

Similarly, the Lennard-Jones (LJ) potential is given by the following expressions:

\[
E_{ij}^{\text{LJ}} = f(r_{ij})V_{ij}^{\text{LJ}}
\]

(3.3a)

\[
V_{ij}^{\text{LJ}} = 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]
\]

(3.3b)

where \(f(r_{ij})\) represents a set of cut-off functions and \(\epsilon_{ij}\) and \(\sigma_{ij}\) are the potential barrier and the finite distance at which the inter-particle potential is zero respectively. Finally,
the torsional terms are given by:

\[ E_{ijkl}^{\text{TORSION}} = f(r_{ij}, r_{jk}, r_{kl}) V_{ijkl}^{\text{TORSION}} \]  \hspace{1cm} (3.4a)

\[ V_{ijkl}^{\text{TORSION}} = \epsilon \left[ \frac{256}{405} \cos^{10} \left( \frac{w_{ijkl}}{2} \right) - \frac{1}{10} \right] \]  \hspace{1cm} (3.4b)

where \( r_{ij} \), \( r_{jk} \), and \( r_{kl} \) are bond lengths, \( w_{ijkl} \) is the dihedral angle, and \( \epsilon \) is the barrier height of the potential. The constants in the above equation are chosen to ensure symmetry of the torsional potential. Exploring the description of the AIREBO potential is important as this work explores the uncertainty propagation through its parameters and the sensitivity of the properties of graphene when using this interatomic potential. For a complete description and formulation of the AIREBO interatomic potential the reader is referred to [66].

### 3.3 Graphene Properties from Molecular Dynamics

Graphene properties are obtained by performing MD simulations using LAMMPS [29] software. Graphene topology is created using Python to arrange the atomic positions of the carbon atoms into a two-dimensional hexagonal lattice and ensuring periodic boundary conditions are met in all directions. For the purpose of this investigation elastic constants, cohesive energy, defects (vacancy, divacancy, bond rotation, and grain boundary) formation energy, thermal conductivity, and strain-stress relationship are selected as observables. Simulations for all the observables are performed on a graphene sheet containing around 2400 or 3600 atoms except for the strain-stress relationship where the graphene sheet contains approximately 7500 atoms to account for size effects.

#### 3.3.1 Elastic Constants

Graphene mechanical properties are better described by its elastic constants. The relationship between stress \( (\sigma_{ij}) \) and strain \( (\epsilon_{kl}) \) tensors can be found using Hooke’s law [67]:

\[ \sigma_{ij} = \sum_{kl} C_{ijkl} \epsilon_{kl} \]  \hspace{1cm} (3.5)

Since graphene is a two-dimensional material, only the in-plane linear elasticity needs to be considered which reduces the constitutive matrix as follows:

\[ C = \begin{pmatrix} C_{11} & C_{12} & 0 \\ C_{12} & C_{11} & 0 \\ 0 & 0 & \frac{1}{2}(C_{11} - C_{12}) \end{pmatrix} \]  \hspace{1cm} (3.6)
In terms of the MD simulations, to calculate graphene elastic constants the initial system is allowed to relax and the stress tensor is calculated. Deformation of the system is then performed in the required direction and the energy is minimized to calculate the new stress tensor. The elastic constants are then calculated by the ratio between the stress and the strain for each direction.

3.3.2 Energetic Properties and Defects

This section discusses how energetic properties such as cohesive and defect formation energies are calculated using MD simulations.

**Cohesive Energy**

The binding or cohesive energy is the energy required to break all bonds associated with the structure. Cohesive energy is calculated as the average atomic energy. The entire simulation is performed at a temperature of 0 K, since results are independent of temperature. The energy per atom is retrieved after the system is allowed to fully relax and become stable.

**Defect Formation Energy**

Defects formation energies, in particular vacancy, divacancy, and bond rotation, are calculated by removing one or two atoms or by modifying the atomic positions of a random atom from the graphene sheet respectively. Energy minimization ensures that the system is stable and that all atoms fall into their corresponding atomic positions. In general, defect formation energy is calculated using the following relationship [68, 69]:

$$E_{sys} = E_{defect} - \frac{N_{defect}}{N_{pristine}} E_{pristine}$$  \hspace{1cm} (3.7)

where $E_{defect}$ is the total energy of the system containing the defect; $N_{defect}$ is the total number of atoms on the defective system; $E_{pristine}$ is the total energy of the system without any defects; and $N_{pristine}$ is the total number of atoms without the defect. Note that for bond rotation the number of atoms in the sheet is unchanged therefore, the final energy of the system is just the difference from energy of the defect and pristine sheets. The different formation energies come from the fact that graphene has the ability of forming non-hexagonal rings in the presence of defects. A single point defect like a C-C bond rotation or Stone-Wales defect yields a structure where four hexagons are transformed into two pentagons and two heptagons (5-5-7-7) configuration. A single
missing lattice atom yields a topology containing a single pentagon and a nine membered ring (5-9) configuration. The final defect included in this paper is the missing of two lattice atoms where the structure re-configures into two pentagons and one octagon (5-8-5). The graphene configurations of the different defects is depicted in Figure 3.1.

**Grain Boundary Formation Energy**

Graphene’s line defect formation energy can also be calculated by a slight modification of the equation 3.7. A comprehensive formulation for grain boundary formation energy is given by Liu [70] as follows:

\[
E_{sys} = \frac{N(E_{defect} - E_{pristine})}{2L_{GB}}
\]  

(3.8)

In this equation, \( N \) represents the total number of atoms in the system, \( L_{GB} \) defines the length of the grain boundary; all other terms have been defined in the previous sections. In order to provide periodic boundary conditions, a bicrystal topology with symmetric tilt boundaries was created (Figure 3.1). The system is heated to a high temperature (\( \approx 700 \) K) at a rate of 5 K/ps and then lowered to room temperature at a rate of 25 K/ps to ensure that the grain boundary becomes stable. The system is then minimized to calculate the formation energy.

Although there are multiple ways a grain boundary can be formed, this paper uses the coincidence site lattice (CSL) theory as described by [71] with a misorientation angle of \( \theta = 9.4 \) deg to form the grain boundary. This yields a configuration of one pentagon and one heptagon along the grain boundary line. This configuration is regarded as \( \Sigma = 37 \), where \( \Sigma \) is an indication of the ratio between the CSL cell and the unit cell for graphene [61, 71].

### 3.3.3 Thermal Conductivity

An additional property of graphene that is measured is the thermal conductivity of graphene. Thermal conductivity in graphene is regarded as one of the most impeccable properties of the material reaching values in the range of 4000 to 6000 W/mK [72] however, since thermal conductivity in graphene depends greatly on the phonon mean free path, it becomes computationally inefficient to analyze it using MD. It is possible; however, to retrieve the value of thermal conductivity as a function of the length of the graphene sheet as given by [73]. Using a sheet of approximately 0.5 \( \mu \)m in length, it is possible to set up one side of the sheet as a hot reservoir, while setting the opposite
3.3.4 Strain-Stress Relationship

Finally, the strain-stress curves are obtained by initially relaxing and minimizing the energy of the large graphene sheet. The temperature is raised to room temperature $\sim 300$ K and tensile deformation is performed to the graphene sheet. The stress tensor is calculated while keeping track of the strain applied at each step to obtain the curves. The sheet is strained at a rate of $0.1\%$ meaning that at each time step in the simulation the sheet length is increased by $10\%$ until failure.
Table 3.1: Reference values for graphene observables used in this thesis

<table>
<thead>
<tr>
<th>Observables</th>
<th>Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{11}$</td>
<td>1150 - 1208.78 GPa</td>
<td>[63, 66, 74]</td>
</tr>
<tr>
<td>$C_{12}$</td>
<td>150 - 208.81 GPa</td>
<td>[63, 66, 74]</td>
</tr>
<tr>
<td>$E_{\text{cohesive}}$</td>
<td>-7.49 ± 1% eV</td>
<td>[75–77]</td>
</tr>
<tr>
<td>$E_{\text{vacancy}}$</td>
<td>7.0 ± 7% eV</td>
<td>[61, 78, 79]</td>
</tr>
<tr>
<td>$E_{\text{divacancy}}$</td>
<td>7.2 ± 5% eV</td>
<td>[61, 78, 79]</td>
</tr>
<tr>
<td>$E_{\text{bond-rotation}}$</td>
<td>5.0 ± 2% eV</td>
<td>[61, 78, 79]</td>
</tr>
<tr>
<td>$E_{\text{grain-boundary}}$</td>
<td>0.223 ± 0.5% eV/A</td>
<td>[71, 80]</td>
</tr>
<tr>
<td>$T_{\text{conductivity}}$</td>
<td>400 ± 10% W/mK</td>
<td>[73]</td>
</tr>
</tbody>
</table>

3.4 Uncertainty Propagation and Quantification

This section describes how parametric variability in the parameters of the AIREBO potential influence the properties of graphene predicted using MD. In order to provide an insightful discussion, Table 3.1 shows a list of the observables and their respective reference values found in the literature. Uncertainty is initially introduced to the LJ portion of the AIREBO interatomic potential followed by the results of introducing variability to the complete interatomic potential. The results are presented by evaluating the sensitivity of the properties of graphene described in the previous section.

3.4.1 Uncertainty Propagation in LJ Potential

Linking MD simulations with the field of uncertainty quantification has led to an increased popularity in the use of the LJ potential. This simple potential presents a fast and effective way to apply uncertainty propagation techniques. This initial study is aimed to confirm that uncertainty in the parameters of the LJ potential do not introduce significant variability in the predicted observables. For this study, only the elastic constants, cohesive energy, and vacancy formation energy are considered as quantities of interest.

Monte Carlo sampling is performed to gain some initial insights into the sensitivity of the quantities of interest when variabilities ranging from ±0.01% to ±5% are imposed in the LJ parameters of the AIREBO potential. In particularly, uncertainty is introduced to $\epsilon_{ij}$ and $\sigma_{ij}$ in equation 3.3b so that they could be sampled from a uniform distribution centred at their nominal value. Figure 3.2 shows the scatter plots of the quantities of interest when a high variability of ±5% is included in the parameters of the LJ potential. Scatter plots for the LJ analysis show a dispersion in the output quantities but even for high variabilities this dispersion is within the limits of the nominal values described in Table 3.1. As part of the AIREBO potential, the LJ term was chosen over other models
Figure 3.2: Scatter plots for (a) $C_{11}$, (b) $C_{12}$ elastic constants, (c) cohesive energy, and (d) vacancy formation energy following a $\pm 5\%$ variability in the LJ parameters of the AIREBO potential.

due to its mathematical convenience and with the expectation of small inaccuracies for many applications [66]. Similarly dispersion and short-range repulsion effects might be more representative in observables that describe the electronic properties of the material. The inaccuracies encountered by this potential term can further be amplified when parametric uncertainty is included in the system and care must be taken when evaluating additional quantities of interest. However, in this study the small dispersion in the data indicates that the terms in the LJ potential do not impact the sensitivity of the observables in a significant way. Further analysis was reasoned as unnecessary for the parameters of the LJ potential alone and the attention could now focus in the collective set of parameters that constitute the AIREBO potential.

### 3.4.2 Uncertainty Propagation in AIREBO Potential

From the functional forms of the AIREBO potential described in equation 3.1, ten parameters were chosen to introduce uncertainty: Five parameters from the REBO potential which include $Q_{CC}$ and $\alpha_{CC}$ from the repulsive term, and $\beta_{CC1}$, $\beta_{CC2}$, and $\beta_{CC3}$ from
the attractive term. Four parameters from the LJ potential including \( r_{CC}^{LJ_{\text{min}}}, r_{CC}^{LJ_{\text{max}}}, \epsilon_{CC}, \) and \( \sigma_{CC}. \) As well as the \( \epsilon_{CCCC}^T \) parameter from the TORSION potential. Similar to the analysis done for the LJ parameters, the initial investigation was performed using Monte Carlo techniques. The same variability range used in the LJ potential analysis (\( \pm 0.01\% \) to \( \pm 5\% \)) was applied. Further investigation of the sample space is performed using sparse grids.

**Monte Carlo Analysis**

The Monte Carlo technique was used to sample the parameters with a sample size of \( N = 500 \) and assuming a uniform distribution where each parameter changed from its nominal value based on the given variability. For instance, for the LJ parameter \( \sigma_{ij} \) with a nominal value of 3.40 Å, its distribution on a \( \pm 5\% \) variability was given by \( U[3.23, 3.57] \). The basic statistics of the results are shown in Figure 3.3. Each boxplot shows how the variability in the potential parameters affects the output quantities. The top and bottom of the box represent the first (Q1) and third (Q3) quartiles respectively, the center line is the median, the top and bottom whiskers represent \( Q1 - 1.5 \times (Q3 - Q1) \) and \( Q3 + 1.5 \times (Q3 - Q1) \) respectively or \( \pm 2.698\sigma \), and the dots correspond to any outliers.

![Graphs showing statistical results](image)

**Figure 3.3:** Statistical results of (a) \( C_{11} \) elastic constant, (b) \( C_{12} \) elastic constant, (c) cohesive energy, and (d) vacancy formation energy when variability is introduced to the AIREBO potential.
For variabilities below and including ±0.5% the predicted observables match those values found in the literature. For the ±1.0% variability, it can be noted that the observables lie within a reasonable range even under the presence of outliers. This is particularly true for formation energies of divacancy, bond rotation, grain boundaries, as well as for thermal conductivity where the range of the outliers is beyond the limits of the observables. As variability increases, it is noticeable that the number of outliers significantly increases and the variability in the MD predictions increases significantly. The star depicted in the plots represent the mean value for each of the observables at the given variability. The shift in the mean provides an indication of how much error is propagated. An interesting fact to note is that the thermal conductivity results are not as spread out as the other observables. This can be accredited partly due to an opposite effect from the LJ and TORSION terms of the interatomic potential as described in [81]; and because in this approach thermal conductivity is dependent on the length of the graphene sheet. The thermal conductivity results displayed in Figure 3.3h show a larger dispersion of values for higher variabilities. However, the mean seems to be constant regardless of the variability used, which might indicate that some effects are not fully
captured when measuring this observable and further studies might be needed to assess this property.

In certain cases where large variability is present, the extreme results can be attributed to a failure in the minimization approach used by the MD simulation. These set of results provided an insight into how the variability on the parameters of the AIREBO potential affect the quantities of interest.

Further investigation showed that for certain samples of interatomic potentials with a high variability of $\pm 2.5\%$ or $\pm 5\%$, it was possible to obtain reasonable values for some observables but not for others. For instance, a sampled potential that could predict the elastic constants within a tolerance level would not be able to predict the vacancy or divacancy formation energy accurately. This motivated the use of better methods to explore the sample space and the introduction of surrogate models as described in the following sections.

**Sparse Grid Analysis**

A more rigorous technique for selecting points from the sample space was used to study the statistics of the MD predictions. By applying the sparse grid technique at three different levels the mean and variance was obtained to study the convergence of the results. In theory, as the level in the sparse quadrature is increased, the statistical moments should reach convergence. For this analysis, only three levels were selected to sample the parameter space, since it becomes computationally expensive to perform simulations at higher levels as it requires more than 850 MD simulations.

Similar to the Monte Carlo technique, variabilities ranging from $\pm 0.01\%$ to $\pm 5\%$ were applied to all the studied parameters in the interatomic potential. However, only the mean and variance for the first three levels at the highest variability of $\pm 5\%$ is shown in Figure 3.4. From the use of sparse grid techniques it was seen that a similar effect to the Monte Carlo simulations was found. At such high variability, some points in the sample space yielded results that were inconsistent with the required results. This means that certain combinations of parameters in the AIREBO potential greatly affect the resolution of quantities of interest.

The motivation behind using sparse grids is to reduce computational cost while being able to sample all the parameter space. In other words, it allows the investigation of the full design space while reducing the number of degrees of freedom and maintaining the overall accuracy. Using a set of four interatomic potentials obtained using the sparse grid sampling method, the stress-strain relationship was also studied to further investigate how parametric uncertainty in the AIREBO potential affect the mechanical properties of
Figure 3.4: Quantities of interest mean and variance when using sparse grid technique to sample parameter space.
the material. Figure 3.5 shows the $\sigma - \epsilon$ curves using the sampled interatomic potentials (dotted line) in comparison with the original curve (solid line).

![Graphene stress-strain curves using four different AIREBO potentials with a ±5% variability.](image)

Figure 3.5: (a) Graphene stress-strain curves using four different AIREBO potentials with a ±5% variability. The solid black line represents the $\sigma - \epsilon$ curve for pristine graphene using the nominal parameters of the interatomic potential. (b) MD tensile test simulation of pristine graphene sheet at breaking point.

The fact that these analyses have shown a large variability in the output results give an indication that care must be taken when using any interatomic potential. As seen from these results, variability in the fitted parameters of the interatomic potential can result in observables that do not converge to their required values.

To fully understand the effect of introducing uncertainty to the parameters of the AIREBO potential, an extensive investigation was performed to rule out any effects arising from the parameters of the MD simulation setup. That is, energy and force tolerance limits were investigated as well as any deviations from the physical topology of the system. After careful adjustment of these parameters, no significant improvements were noticed suggesting that the extreme results were produced by the variability in the potential parameters alone, thus motivating the use of surrogate models to perform a sensitivity analysis.

### 3.5 Surrogate Model Applied to Graphene

In this study, MD is regarded as a black-box where a set of interatomic potentials parameters are considered to be the inputs and the properties predicted are treated as output quantities. Simulations of this nature tend to be computationally expensive and take a
significant amount of time even when the runs are parallelized. These issues combined with the fact that MD algorithms greatly depend on the interatomic potential being used, motivated the use of surrogate models in order to perform an exhaustive uncertainty analysis. This section describes the creation of the surrogate model using the existing data obtained from the sampling performed in the previous section. Similarly, a complete description of the model validation is presented.

### 3.5.1 Bayesian Gaussian Process Model

Using a statistical framework based on the Bayesian Gaussian process approach a model that allows the prediction of the quantities of interest without the need to run MD simulations was constructed.

**Leave-one-out Validation**

The initial step in generating an accurate model is to separate an existing dataset into a training and a testing set; where the surrogate model aims to make accurate predictions on the testing set based on the tuning done with the training set. The Monte Carlo dataset from the earlier study with $\pm 5\%$ variability in the input parameters is used as the training data to build the Gaussian process model. The leave-one-out approach of the model is presented in Figure 3.6. This approach is used to validate the model and show its ability to predict the quantities of interest.

Using the leave-one-out approach is not an optimal method for determining the validity of the model as it tends to give estimations of the prediction error that are more variable than other methods. However, the LOO approach provides a general intuition on the performance of the metamodel. The LpO cross-validation method is also applied by leaving $p = 20$ training points out at each iteration. The purpose of this validation method is to understand the error produced by the metamodel when estimating several unseen points instead of only one. This method also provides an estimation of the minimum number of samples required for the training of the model to give a good estimation of the hyperparameters. Figure 3.7 shows the LpO method applied to four of the quantities of interest in order to provide a comparison to the LOO approach.

It can be observed from Figures 3.6 and 3.7 that the LpO gives larger values of error compared to the LOO approach. This is particularly true for defect formation energies such as vacancies, divacancies, and grain boundary formation. By noticing the differences of the two approaches it is safe to assume that the model requires a large number of training samples to obtain a good approximation.
Figure 3.6: Leave-one-out validation approach of the surrogate model for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacancy}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}}$. 
Chapter 3. 2D Material System: Graphene

Figure 3.7: Leave-p-out validation approach for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{divacancy}}$, and (d) $E_{\text{grain-boundary}}$ formation energies. Validation obtained using $p = 20$ training points.

Standardized Cross-Validation Residuals

The SCVR is described in Chapter 2 as the ratio between the true residuals and the true standard deviation of the residuals. This values give an estimation of the goodness of the model by assuming that if the Gaussian process model is appropriate the SCVR will lie in the interval $[-3, +3]$. This implies that the predictions of the model will lie within $\pm 3\sigma$ limits with a high level of confidence. Figure 3.8 shows the plots of the SCVR for all the observables.

The Bayesian GPM is created based on the realizations made when sampling the parameter space of the AIREBO potential. It can be noted from the plots that predictive capabilities of the model are generally adequate and error bars seem to be tight. This model can now be used to perform a sensitivity analysis and understand which parameters influence the selected observables.
Figure 3.8: Standardize cross-validation residuals for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacancy}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}} \pm 3\sigma$ are marked with dotted lines.
3.5.2 Model Comparison: GPMs and RBFs

A surrogate model based on radial basis functions with Gaussian kernel was also developed to estimate the output variables based on the sampled data. This section provides an insightful comparison between the Gaussian process surrogate and the RBF metamodel. To provide an impartial comparison, both models are validated using a k-fold approach, in which the data is partitioned in “k” smaller sets and k − 1 sets are used to train the model while the remaining part of the data is used for validation. For this comparison, and to measure the robustness of the model, a k-fold approach with k = 2 is used. That is, the data is divided in two sets of N = 250 points each for training and testing purposes. Figure 3.9 shows a comparison of the two model for six of the quantities of interest.

Even though the expressions for the predictions of the GPM and the RBF models are similar, the main difference between the models is the estimation of the hyperparameter \( \theta \) described in equations 2.23 and 2.28 for the GMP and RBF models respectively. Although these models share a similar structure their results are very different as shown in Figure 3.9, where it can be noted that the RBF surrogate is less accurate in predicting the observables as compared to the GPM model. The flexibility of the GPM model also provides a method of calculating error bars since the model produces an estimate of the uncertainty involved in making predictions at unseen points by calculating the posterior variance. Computationally the flexibility of the GPM surrogate implies a larger time to make the required predictions because the shape parameter needs to be optimized. In average the training and testing of the model takes \( t_{RBF} = 0.241598 \) s and \( t_{GPM} = 0.869938 \) s for the RBF and GPM surrogates respectively. Even though the GPM model requires a computational time that is almost four times larger than the RBF model, it is selected for the rest of this analysis for the flexibility it offers in proving an error estimation of the prediction.

3.6 Sensitivity Analysis

This section makes use of the GPM surrogate model to compute the main effects, the sensitivity indices, and the interaction effects of the interatomic potential parameters.

3.6.1 Main Effects

Main effects are obtained by integrating the contribution of each variable from the input parameters of the AIREBO potential. For a variability of \( \pm 5\% \) the effects of each input
Chapter 3. 2D Material System: Graphene

Figure 3.9: GPM and RBF model comparison for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{bond-rotation}}$, and (f) $E_{\text{grain-boundary}}$ formation energies. A k-fold cross validation is used with $N = 250$ points for the training and testing sets.
variable to the $C_{11}$ elastic constant, vacancy formation energy, cohesive energy, and grain boundary formation energy are depicted in Figures 3.10 and 3.11. These observables were selected as examples to showcase the main effects of input parameters.

The main effect plots show some interesting results. First, it can be seen that parameters from the LJ term (i.e. $r_{LJ}^{CC}$, $\rho_{LJ}^{CC}$, $\epsilon_{CC}$, and $\sigma_{CC}$) produce flat lines for the vacancy and grain boundary formation energies, while it produces a linear effect for the elastic constant and the cohesive energy. Since the effects of these parameters are not significant, this confirms that the LJ term in the AIREBO potential does not contribute to a large effect to the description of these observables. Second, the torsional term in the potential ($\epsilon_{CCC}^T$) shows a minimal effect on the observables. However, it is interesting to note that the confidence interval for the grain boundary formation energy shows a non-linear effect, suggesting a more significant effect to this observable. This can be explained due to the small torsion the graphene sheet encounters due to the nature of the grain boundary. Finally, it can be seen how non-linear terms produce a large effect on the output results, in particular $\alpha_{CC}$ and $\beta_{CC1}$ are the parameters with largest influence on the presented observables. From the analysis it was found that in general these two parameters affect in large extent any of the observables that were reproduced. The parameters from the attractive and repulsive terms of the AIREBO potential determine the conditions for which graphene structure and properties are characterized since they affect the strong atomic bonding of the material. In general, the sp$^2$ hybridization of carbon atoms in graphene plays an important role since the effects of the attractive and repulsive terms are scaled based on the bond order [66]. It should also be pointed out that no quantum effects are present in the description of the AIREBO potential therefore; the characterization of these terms is not fully accurate.

### 3.6.2 Sensitivity Indexes

The sensitivity indices give an indication on the variance of the input parameters. Similar to the main effects, the sensitivity indices quantify the influence of the input parameters to the selected observables. Figure 3.12 depicts the effects of each parameter to the quantities of interest. From the pie-charts it is possible to see that there is a dissimilar effect from the all input parameters; in particular there is a high effect is shown from the $\alpha_{CC}$ and $\beta_{CC1}$ inputs. In general this large influence is evidence of a high non-linearity from the input parameters. It is interesting to note that defect formation energies, in particular vacancy, divacancy, and grain boundary formation energy tend to have a similar effect from their input parameters which can be attributed to the fact that these
Figure 3.10: Main effects for $C_{11}$ and $E_{\text{vacancy}}$ formation energy. Abscissa shows a normalized value for each input parameter in the AIREBO potential. Red dotted line signifies a ±95% confidence interval based on the standard error.
(a) Cohesive energy

(b) $E_{\text{grain-boundary}}$ formation energy

Figure 3.11: Main effects for $E_{\text{cohesive}}$ and $E_{\text{grain-boundary}}$ formation energy. Abscissa shows a normalized value for each input parameter in the AIREBO potential. Red dotted line signifies a ±95% confidence interval based on the standard error.
Figure 3.12: Pie charts showing the sensitivity of the parameters in the AIREBO potential for (a) $C_{11}$, (b) $C_{12}$, elastic constants (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{divacancy}}$, (f) $E_{\text{bond-rotation}}$, (g) $E_{\text{grain-boundary}}$ formation energies, and (h) $T_{\text{conductivity}}$. 
energetic properties are formed in a similar fashion, thus allowing a restructure of the graphene sheet into more stable configurations consisting of pentagons and heptagons.

Another interesting effect described by the charts is the difference effect that is shown from the $\alpha_{CC}$ and $\beta_{CC_1}$ inputs to observables that are obtained from the same process such as $C_{11}$ and $C_{12}$. In the transversal elastic constant there is a much larger effect from the parameters of the repulsive parameters than the attractive terms. This can be explained as the AIREBO potential includes bond properties to fully describe the elasticity of the material.

### 3.6.3 Interaction Effects

To understand non-linear effects, it is possible to construct the interaction effects from two variables to the output values. To provide a graphical description of this method, the input parameters are selected from a general normalized grid and the response of the metamodel is observed in a contour plot. The study of such contour plot can provide an estimation of the functional relationship between the two variables to the output quantity. Figure 3.13 shows the contour plot representing some of the interactions between input variables; the figure displays four different functional forms in which the interaction effects happen to describe the elastic constant $C_{11}$.

Figure 3.13a shows the interaction between a parameter from the repulsive term and a parameter from the LJ potential. From the plot it can be seen that the functional form is parabolic for the LJ term. That is, as the repulsive term increases in value, two values of the LJ term can be used to quantify the observable in the same way. Similarly, Figure 3.13b shows a linear response between the two variables from the REBO term in the potential. This is expected as these terms work together to define the attraction of particles in the system. Figure 3.13c, shows an interesting effect from the parameters of the LJ potential. This effect can be attributed to the fact that the $r_{CC_{LJmin}}^C$ term is used as a switching term for the rest of the terms of the LJ potential meaning that after a certain cutoff value, the $\sigma_{CC}$ term returns from a maximum value to a minimum value. Finally, the effect shown in Figure 3.13c is very flat, which shows that there is almost no interaction between the variables.
Figure 3.13: Interaction effects from different input variables to describe the response of the $C_{11}$ elastic constant.
Chapter 4

3D Material Systems: Metals

This chapter applies the methodology described in Chapter 2 to a metallic system, in particular, aluminium. As in the previous chapter, the topology of the material is initially presented followed by a definition of the interatomic potential used when performing MD simulations. Finally, numerical studies are carried out to investigate how uncertainties in the parameters of the interatomic potential influence the predictions made using MD.

4.1 Aluminium Topology

Metallic systems exist everywhere and are popular material systems used in aerospace applications. Understanding their mechanical and thermal properties is crucial for the development of good structures. However, it is important to realize that the mechanical, electrical, or thermal behaviour of the material arises from its microstructure. The science that studies the atomic arrangement in three-dimensional material systems is known as crystallography and it is based on experimental observations and techniques to understand the material structure. Crystallography divides materials into crystal systems or groups in order to classify them. Most metallic systems belong to the cubic crystal system where the unit cell is in the shape of a cube. Further classification is based on atomic packing inside the unit cell, where crystals are classified as simple cubic, body centred cubic (BCC) or face centred cubic (FCC) [82]. The importance of crystallographic becomes evident in the description of defects.

The focus of this chapter is on aluminium, a material system that contains atoms arranged in a face centred cubic structure (See Figure 4.1). This topology is easily generated when performing MD simulations as the lattice vectors are described by orthogonal basis and scaled by the lattice constant of 4.050Å. Understanding the structural arrangement of atoms in metallic systems provides an idea of some properties of the material. For
instance, FCC structures contain atoms that are closely packed in comparison to BCC structures, this allows the system to allow for more plastic deformation as atomic planes can easily slide by each other, affecting the material ductility [83]. Since FCC structures define the bulk properties of the metallic system of interest, the sections that follow will refer to FCC and aluminium systems interchangeably. The present methodology could be also applied to other metallic systems of the same structure such as nickel, copper or silver.

4.1.1 Defects in FCC Structures

Defect analysis is very important in defining the properties of the material. Perfect materials are impossible to find in nature; in fact, if metallic systems were composed with perfect crystalline structures their properties would be restricted. Furthermore, tailoring metal properties to specific values would be a challenging task. Defects in crystal structures are classified by their dimensionality, that is, a point defect such as vacancy or interstitial defects are considered zero dimensional, dislocations are considered one dimensional, and planar defects such as grain boundaries or stacking faults are considered two dimensional defects [84].

Point defects modify a single lattice point either by removing an existing atom in the case of a vacancy defect, or by introducing impurities around the lattice point. These defects usually occur during the solidification of the material either due to vibration or rearrangement of atoms or by the addition of similar or smaller size atoms in the structure [85]. Point defects and their effects on the material properties are well understood as they change the energetic structure of the system. Vacancies, for instance, displace atoms
around and create a small tensile stress in the vicinity [86]. Interstitial defects, on the other hand, produce some interesting effects in the softening and glass formation abilities of the material [87].

Dislocations are line defects where atoms in the structure are out of position and are generated when stress is applied to the material. Dislocations affect the plasticity of the material, in particular, affecting the shear strength of the material by allowing bonds to break and reform. Dislocations play an important role when studying fracture, fatigue, and creep of materials because they allow the study of crack initiation and displacement.

Surface or planar defects are of increasing interest in defining the properties of crystalline materials as they produce dramatic changes in their structure. The most common planar defects in crystalline structures are stacking faults and grain boundaries. The former is created by an interruption in the atomic stacking sequence of atoms in closed packed structures. For FCC structures the stacking sequence is generally “...ABCABC...”; where A, B, and C refer to a distinct atomic layer position. Stacking fault occurs when the sequence is altered in any order causing a change in the energy of the system [83]. The major influence of stacking faults is the ability of the material to alloy with other metals.

A different planar defect is the grain boundary or interface between two grains of a polycrystalline material which tends to affect not only the mechanical, but also the electrical and thermal properties of the material. Grain boundaries reveal their importance during plastic deformation in polycrystalline materials because many dislocations pile up along the grain boundaries [7] and also due to effects produced by grain boundary sliding [88]. Ongoing research on this field has also shown particular interest in the analysis of grain size as it seems to be the most contributing factor to limit the ductility of nanostructures. Grain boundaries and sizes are also used to characterize and enhance the strain hardening of materials by introducing a bimodal grain size distribution [89]. That is, when two different sizes of grains predominate in the mean grain size distribution of the material. The idea behind this method is that the larger grain size will provide more ductility while the smaller grain size will provide the strength. This is because grain boundaries prevent dislocations motion inside the grains.

Not so much a defect, but a property of metallic systems that is important to mention is their capability of creating free surfaces. This can be thought as the new surfaces that get generated when a bulk material is divided in two pieces. A free surface in metallic systems allows the bonding to other materials and affects the shape of the material. In general, free surfaces are derived from the pattern of atomic packing on the crystal plane that lies parallel to the surface and its behaviour can also be inferred from this
pattern. In this work, most of the aforementioned defects are analyzed in the sense of MD simulations and uncertainty analysis.

4.2 EAM Interatomic Potential

The fact there are numerous interatomic potentials available to describe a single material system illustrates the weakness of MD and its dependence on interatomic potentials. For the case of aluminium, a simple search shows that it is possible to find more than seven potentials capable of describing the atomic interactions of the material. This work will employ the Embedded Atom Method (EAM) as constructed by Mishin and Farkas [90] to perform MD simulations of aluminium. This section aims to provide a general description of the potential and its importance to simulating this mono-atomic material.

The EAM interatomic potential was originally developed in 1983 by Daw and Baskes [91] to understand metal impurities; in particular the effects on the mechanical properties of metals when hydrogen atoms are embedded into the structure. The model was created to overcome the difficulties in predicting metal properties when using simple pair-wise interatomic potentials. The idea behind the EAM potential is to use the concept of local density and the quasiatom theory where the “embedded” atom experiences an uniform environment or effect from all other atoms. To get an overall description of the atomic interactions, the model treats every atom as an “embedded” atom and considers the energy of the system as a functional of the unperturbed electron density and a function of the embedded atom type \((Z)\) and position \((R)\) as given by the following expression:

\[
E = F_{Z,R}[\rho_h(r)]
\]  

(4.1)

where \(\rho_h(r)\) is the system’s electron density when unperturbed, the functional \(F\) is unknown and rather complicated; however, it can be approximated using the quasiatom theory such that the total energy is a function of the embedding energy \(F\); that is:

\[
E_{total} = \sum_i F_i(\rho_h(r_i))
\]  

(4.2)

The embedding function is defined as the energy of the embedded atom in a uniform electron gas relative to the atom separated from the electron gas [91]. However, by assuming equation 4.2 describes the total energy of the system, short-range interactions such as the core-core repulsion between atoms are neglected. A more complete formulation is
created by adding a pair potential function, such that the total energy is calculated by:

\[
E_{\text{total}} = \sum_i F(\bar{\rho}_i) + \frac{1}{2} \sum_{i,j} \phi_{ij}(r_{ij}) \tag{4.3a}
\]

\[
\bar{\rho}_i = \sum_j \rho(r_{ij}) \tag{4.3b}
\]

where \(\bar{\rho}_i\) is the atomic electron density induced at atom \(i\) by all other atoms in the system and \(\rho\) is defined as the local atomic density function associated with the distance \(r\) between atoms \(i\) and \(j\). The function \(\phi\) is the pair potential as a function of the atomic distance. Note that the EAM potential does not account for three or more body terms which makes it difficult to account for covalent effects. Equation 4.3 is known as the generalized form of the EAM potential and proves to be useful in describing defective metallic systems.

One of the challenges in developing the EAM potential is determining the embedding and pair-potential functions. In the case of aluminium, Mishin et al. [90] use a semi-empirical approach where they combine \textit{ab initio} calculations with experimental results to define these functions. The functions are obtained by creating a database containing several material properties and using a weighted fitting procedure to estimate the functional forms of equation 4.3. However, when a particular property is not included in the fitting procedure, the potential will not be able to accurately approximate that property. Figure 4.2 depicts the three functions that describe the interatomic potential for aluminium. The reader is referred to [91] and [90] for a complete description of the EAM theory and the fitting procedure applied to the aluminium potential.

![Figure 4.2: Individual functional forms of the aluminium’s EAM potential.](image)
4.3 Aluminium Properties from Molecular Dynamics

Similar to the previous chapter, aluminium material properties are obtained using LAMMPS software to perform the MD simulations. The topology of the material is created using definitions available in the LAMMPS code. For instance, the FCC structure can be specified in the input file using the *lattice* command. Similarly, several conditions such as orientation, size, and shape can be detailed to construct the structure containing defects. The properties being discussed in this chapter are: elastic constants, cohesive energy, lattice parameter, and formation energies for vacancies, interstitials, stacking faults, grain boundaries, and free surfaces. Each property is obtained using the minimum number of atoms allowed in the MD simulation in order to minimize computational resources.

4.3.1 Elastic Constants

The overall description of elastic constants is given in equation 3.5. In general, this equation should determine 81 independent stiffness constants; however, due to material symmetry this number is reduced. For instance, Voigt symmetry reduces the number to 21 constants. Further reduction is obtained due to the crystallographic symmetry of the material; in particular, for cubic structures the material is characterized with three constants: Longitudinal compression is determined by $C_{11}$, transverse expansion is given by $C_{12}$, and shear modulus is obtained from $C_{44}$ [92]. With this in mind, the elastic constant matrix can be written as follows:

$$
C = \begin{pmatrix}
C_{11} & C_{12} & C_{12} & 0 \\
C_{12} & C_{11} & C_{12} & 0 \\
C_{12} & C_{12} & C_{11} & C_{44} \\
0 & 0 & C_{44} & 0 \\
0 & 0 & 0 & C_{44}
\end{pmatrix}
$$

(4.4)

To obtain the above matrix, longitudinal or transverse strains are calculated based on the displacement taken from one particular direction, while shear strain is calculated based on the average displacement taken from two directions. Important properties such as elastic, bulk or shear modulus are also readily available from the elastic constants. From the MD perspective, the system is initially relaxed to ensure the atoms are in a state of minimum energy. The stress is calculated and the system is deformed in six
different directions to ensure that strain is applied in all possible orthogonal directions and combinations. The new stress is calculated and the elastic constants are computed using Hooke's law as the ratio between the stress difference and the strain applied in each direction.

4.3.2 Cohesive Energy and Lattice Parameter

Cohesive energy and the lattice parameter are calculated to ensure that the system simulates the basic parameters of aluminium. For crystal structures, the cohesive energy is the energy of arranging atoms in the crystalline state, while the lattice parameter defines the physical dimension of the unit cell for the FCC structure. By ensuring these parameters are correct it is certain that the simulations are modelling the structure of aluminium. The calculation of these two properties is achieved by minimizing the energy of the system and calculating the average atomic energy to get the cohesive energy, and the ratio between the final and initial lengths of the simulation box length to obtain the lattice constant.

4.3.3 Point Defects Energetic Properties

This section discusses the MD simulations to obtain formation energies of point defects in the aluminium structure.

Vacancy Formation Energy

Vacancy formation energy is defined as the energy of the system when an atom is missing from the structure. The formation energy of a single vacancy is calculated using the following formula:

$$E_{sys} = E_{defect} - \frac{N - 1}{N} E_{pristine}$$  \hspace{1cm} (4.5)

where $N$ is the total number of atoms in the system without defects and $E_{defect}$ is the energy of the system containing the vacancy. In order to calculate this property using MD simulations, the system is first relaxed and the cohesive energy is calculated. A single atom is then removed (See Figure 4.3) from the middle of the structure and the energy is minimized again. Applying the above equation, the vacancy formation energy is computed.
Interstitial Formation Energy

Interstitial formation energy is defined as the energy of the system when one or more aluminium atoms are introduced in the structure. For the addition of a single atom, the energy can be calculated by modifying equation 4.5 as follows:

$$E_{\text{sys}} = E_{\text{defect}} - \frac{N + 1}{N} E_{\text{pristine}}$$  \hspace{1cm} (4.6)

The procedure to calculate the energy is similar to that of the vacancy formation energy. The system is initially relaxed, the atom is then introduced in the system, and the final energy is obtained to perform the calculation. Figure 4.3 shows the addition of the atom in the FCC structure.

![Vacancy and interstitial defects in a FCC structure.](image)

Figure 4.3: Vacancy and interstitial defects in a FCC structure. The colours represent the energy map of the system in [eV].

4.3.4 Energetic Properties of Planar Defects

This work investigates planar defects due to the valuable information that they provide on the properties of the material. This section provides the MD structure to calculate energetic properties of planar defects found in aluminium systems.

Stacking Fault Energy

Stacking faults are very common in close packed structures and they become important in determining the wear resistance of the metallic system as they modify the ability of dislocations to move along the structure. Stacking fault energy is defined as the energy of the system when a stacking fault is present. It is measured per unit area by the following
equation:

\[ \gamma_{SFE} = \frac{E_{\text{defect}} - E_{\text{pristine}}}{A} \]  

(4.7)

where \( \gamma_{SFE} \) represent the stacking fault energy, \( E_{\text{defect}} \) and \( E_{\text{pristine}} \) are the potential energy of the system after and before the defect is formed respectively, and \( A \) is the area of the stacking fault. To produce a stacking fault in MD simulations, the system is divided in an upper portion that is moved slightly with respect to the lower portion until the stacking sequence is broken. The energy is then obtained and the system energy is computed. In order to facilitate the stacking fault creation process, the lattice of the crystal structure is defined in a 111 plane. Figure 4.4 shows the stacking sequence broken down to create the stacking fault.

**Grain Boundary Energy**

Grain boundary formation energy can be calculated using equation 4.7 where the area \( A \) represents the grain boundary area. The construction of the grain boundary for use in MD simulations is done in a similar fashion as the stacking fault. The simulation box is divided in two parts by a plane of interest. Both parts are then rotated by \( \theta/2 \) in a clockwise direction for the upper part and counter clockwise direction for the lower part. Atoms that overlap are then deleted and the grain boundary is created by minimizing the energy of the system. In this work, the grain boundary is created in the 310 plane at an angle of \( \theta = 36.87 \) deg which yields a configuration called \( \Sigma = 5 \) [93]. Figure 4.4 shows the grain boundary configuration used in this analysis.

![Figure 4.4: Stacking fault and grain boundary formation of a FCC structure. The colours represent the energy map of the system in [eV].](image-url)
Table 4.1: Reference values for aluminium observables used in this thesis

<table>
<thead>
<tr>
<th>Observables</th>
<th>Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{11}$</td>
<td>113.8 GPa</td>
<td>[90, 94]</td>
</tr>
<tr>
<td>$C_{12}$</td>
<td>61.55 GPa</td>
<td>[90, 94]</td>
</tr>
<tr>
<td>$C_{44}$</td>
<td>31.59 GPa</td>
<td>[90, 94]</td>
</tr>
<tr>
<td>Lattice</td>
<td>4.050 Å</td>
<td>[90]</td>
</tr>
<tr>
<td>$E_{\text{cohesive}}$</td>
<td>-3.36 eV</td>
<td>[90]</td>
</tr>
<tr>
<td>$E_{\text{vacancy}}$</td>
<td>0.675 eV</td>
<td>[68, 90]</td>
</tr>
<tr>
<td>$E_{\text{interstitial}}$</td>
<td>2.789 eV</td>
<td>[90]</td>
</tr>
<tr>
<td>$E_{\text{stacking-fault}}$</td>
<td>145.2 mJ/m$^2$</td>
<td>[90]</td>
</tr>
<tr>
<td>$E_{\text{grain-boundary}}$</td>
<td>494.0 mJ/m$^2$</td>
<td>[90, 93]</td>
</tr>
<tr>
<td>$E_{\text{free-surface}}$</td>
<td>943.6 mJ/m$^2$</td>
<td>[90]</td>
</tr>
</tbody>
</table>

**Free Surface Energy**

Creation of the free surface is a relatively simple task in MD simulations. The surface is created by making the simulation box larger than the region containing the atoms of the structure and filling the rest of the simulation region with a vacuum. The energy is then calculated by the ratio between the energy difference with and without the vacuum and the area of the surface. The simplest surface to create are those whose normal vectors lies in the 100, 010 or 001 directions.

**4.4 Uncertainty Propagation and Quantification**

Uncertainty is introduced in the EAM potential to study how variability in the interatomic potential parameters affect the predictions made using MD. Table 4.1 provides a list of the observables and their values obtained the MD simulation procedure described in the previous section. Introducing uncertainty in the EAM potential requires a different procedure from other types of potential. The issue in hand is that the parameters from the functional forms of the potential are not readily available. In fact, the EAM potential used for MD simulations is described by a set of tabulated points that describe each of the three functions described in equation 4.3. The EAM potential used to describe the aluminium material system contains 10000 points for each of the functions which makes it computationally inefficient to introduce uncertainty to all of these points. In this analysis, a fitting procedure using splines is performed to obtain an estimation of the functional forms. The coefficients of the interpolation are then used as input parameters for the uncertainty analysis.

Spline interpolation is carried out using a Python code and the number of “knots”
used are defined in such a way that the error between the spline and the original set of points is minimized. Uncertainty is introduced to the embedding function only, to isolate and capture the effects of this function on the properties of the material predicted using MD simulation. This is done for two reasons: first, to generalize the methodology in how uncertainty is introduced; i.e., the same approach can be followed to introduce uncertainty to the other two functions. Second, to avoid the “curse of dimensionality” as fewer interpolation coefficients are required to perform the uncertainty analysis. A caveat to this approach is that it is not possible to determine the combined effect from all the functions on the observables. Figure 4.5 shows the interpolation results of the embedding function as well as the associated interpolation error as a function of the electron density function. A negative error in the graph shows that the interpolation curve overestimates the value of the function at that point. The interpolation method was further analyzed by performing MD simulations to measure the error introduced in the output variables when following this approach. The observable that was most affected by the interpolation method was the elastic constant measuring the transverse expansion \( C_{12} \) which presented a 0.12% error as compared to its original value.

Introducing uncertainty to the coefficients of the interpolation was attained using the Monte Carlo technique using variabilities ranging from ±0.01% to ±5%. A sample size of \( N = 500 \) was used for each of the variabilities and a uniform distribution for each

![Figure 4.5](image_url)

Figure 4.5: Comparison between original points and interpolation result for the embedding function in the EAM potential. The lower plot describes the interpolation error.
of the coefficients was assumed. The statistical distribution is created such that the lower bound is given by the nominal value of the input parameter minus the variability; while the upper bound is given by the nominal value plus the variability. That is, for a variability of \( \pm 1\% \), the input parameters are sampled from a uniform distribution that ranges from \([x_i - 1\%, x_i + 1\%]\), where \(x_i\) represents the nominal value of the \(i^{th}\) coefficient. Figure 4.6 shows the effects of propagating uncertainty to the quantities of interest.

![Figure 4.6](image-url)

Figure 4.6: Boxplots showing the effects on the quantities of interest when uncertainty is introduced at different variabilities to the interpolation coefficients of the EAM potential. Range taken by the observables is shown for clarity.
Figure 4.6: (cont’d) Boxplots showing the effects on the quantities of interest when uncertainty is introduced at different variabilities to the interpolation coefficients of the EAM potential. Range taken by the observables is shown for clarity.
By inspecting the boxplots shown above it can be noted that a small variability in the inputs do not greatly affect the observables. Large variabilities, in particular $\pm 2.5\%$ and $\pm 5\%$, constrain the observables to a range outside their nominal value. Visually these plots look similar, but after careful investigation it was noted that all observables are affected in a different measure. For instance, a variability of $\pm 5\%$ produces a 13.6\% deviation from the mean of the $C_{11}$ elastic constant, a 25.3\% deviation from the mean of $C_{12}$, but a very small deviation of 0.25\% from the mean of $C_{44}$. The largest deviation found at the $\pm 5\%$ variability corresponds to the cohesive energy at 43\%; while the smallest deviation was for the lattice parameter at 0.01\%.

Although the fitting procedure employed by [90] puts a high weight on parameters such as cohesive energy and elastic constants, the introduction of uncertainty limits the performance of the interatomic potential in describing some of these parameters. By conducting this analysis it was possible to show the robustness of the potential and the effects on the output variables when variability is introduced in the fitting procedure. The reader should note that small variability introduced in the input parameters tends to amplify the error in the output quantities of interest.

### 4.5 Surrogate Model Applied to Aluminium

Similar to the study described in Chapter 3, MD simulations are treated as a black-box, the interpolation coefficients are regarded as input parameters and the observables are considered the output variables. Performing MD simulations to obtain several material properties is a computationally exhaustive task since different configurations are required for each observable. This issue motivates the creation and application of the surrogate models described in this section.

#### 4.5.1 Bayesian Gaussian Process Model

This section describes the validation and application of the Bayesian Gaussian process model to MD simulations of aluminium material systems. From an algorithm point of view, the model is built by providing a training dataset derived from the MD simulations. The training set is used to optimize the model by adjusting the required hyperparameters using a maximum likelihood estimator (See Section 2.3.2 for a mathematical description).
Leave-one-out Validation

By performing Monte Carlo sampling, it is possible to construct a dataset that is used to train and test the surrogate model. However, a crucial part of the design is to ensure that the model performs as required. In order to validate the model the leave-one-out approach is implemented to verify the response of the model at evaluating a single unseen point. Figure 4.7 shows the results of applying the LOO method for some of the observables at variability of ±5%; in particular, those output variables that showed a high error as described in the previous section. The model produces a linear trend for all the observables which suggest that the model is able to make accurate predictions of the unseen data point. This is true even when the sampled data is randomized prior to constructing the surrogate to avoid any bias. As mentioned in Section 2.3.5, the LOO approach tends to overestimate the error in the predictions; however, it was found that the surrogate was robust enough to produce a linear response when comparing actual and predicted quantities with both, the LpO and the k-fold approaches.

Standardized Cross-Validation Residuals

The robustness of the model motivated the use of the standardized cross-validation residuals as a method of ensuring the model predicts values within an appropriate confidence interval. Figure 4.8 shows the SCVR for the same observables presented in the previous section. The SCVR corresponds to the number of standard errors that the actual value is above or below the predicted value. This indicates that in some cases the error is actually larger than predicted by the LOO approach. For instance, Figure 4.8a shows several points outside but close to the confidence interval while for the other observables show only a single point outside the confidence interval. The SCVR plots also present information about the bias in the predictions of the model. For example, if an increase or decrease of the SCVR is seen as a function of the predicted points, then a systematic bias exists in the predictions. For the case presented in this analysis, there is no obvious trend of the SCVR which suggest that no bias is presence in the predictions of the model.

By evaluating this metric in combination to the LOO approach, is can be seen that the metamodel is robust enough to make predictions for the quantities of interest. For this particular application, it was noted that even using a conservative number of data points in the training set the model is able to predict the output values accurately.
Chapter 4. 3D Material Systems: Metals

Figure 4.7: Leave-one-out validation approach of the surrogate model for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of $\pm 5\%$ in the EAM potential for aluminium material system.
Figure 4.8: Standardized cross-validation residuals for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{cohesive}}$, (d) $E_{\text{vacancy}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of ±5%. Dotted line represents the confident interval as ±3σ.
4.6 Sensitivity Analysis

Using the Bayesian Gaussian process model applied to the aluminium material system, this section aims to provide insights on the sensitivity of the observable quantities when uncertainty is introduced in the interpolation coefficients of the EAM potential. In this section, the main effects and sensitivity indices are presented.

4.6.1 Main Effects and Sensitivity Indexes

The main effects are computed to understand the effects of each of the input variables on the quantities of interest. They provide an assessment on the importance of each of the variables and how each input parameter influences the results of the metamodel. Although the propagation of uncertainty to the EAM interatomic potential is described in Section 4.4; it is important to mention that the spline interpolation resulted in ten parameters to be used as input variables. Main effects serve as a visualization tool to determine if the input affects the output in a linear, polynomial, or any other functional way. Figure 4.9 shows the main effects for the grain boundary formation energy of aluminium, showing the influence of all variables to the observable.

There are two important things to note from Figure 4.9: first, all variables except $X_1$ and $X_5$ seem to influence the output variable in the same way. These variables show a non-linear effect to the grain boundary formation energy. Second, the influence of variables $X_1$ and $X_5$ is minimal compared all other variables. In general these inputs do not greatly affect the observable as they produce flat lines within their confidence interval. The confidence interval is based on standard error and it is presented to show the full effect these variables have on the output variable.

To further quantify the effects of the input variables on the observables the sensitivity indices are computed. These values are used to further condense the metrics already provided by the main effects. A comparison between sensitivity indices of the grain boundary formation energy and the $C_{11}$ elastic constant are shown in Figure 4.10. From this graph, it can be noted that almost all the input variables provide a similar effect to the predicted quantities. A significant difference between the $C_{11}$ and the $E_{\text{grain-boundary}}$ charts is the effect produced by the fifth interpolation coefficient. In the case of the elastic constant, this input produces a larger and non-linear effect as compared to the irrelevant effect caused on the grain boundary formation energy. By the same token, it should be pointed out that the first interpolation coefficient does not provide a significant effect to any of the quantities of interest which might be attributed to the fact that the embedding function does not account for interaction of atoms that are positioned far
Figure 4.9: Main effects for $E_{\text{grain-boundary}}$ formation energy when using a variability of ±5%. $X_i$ signifies the $i^{th}$ interpolation coefficient used as input parameter.

4.7 Reduced Order Modelling and RBF

The results obtained from the sensitivity analysis showed that most input variables produce a similar effect on the output properties of aluminium. This motivated the creation of a reduced order modelling technique combined with the RBF surrogates. The idea is to reduce the dimensionality of the input space without losing their effects on the output properties of the material. In order to achieve this, the Principal Component Analysis (PCA) technique described in section 2.3.4 is employed.

This technique was applied when combining the interpolation parameters from the
embedding and the pair potential function as described in equation 4.3. By combining these two function, the interpolation method used yielded a set of 35 input parameters. From a dimensionality point of view, creating and validating a surrogate with a large number of inputs becomes a computationally expensive process favouring the use of actual MD simulations. Figure 4.11 shows the results of applying PCA in combination with RBFs for some of the aluminium observables. Note that the predictions of the model are accurate with an error close to zero in most cases. The results shown in Figure 4.11 were constructed using a validation approach with 100 training points and 400 testing points selected randomly from the input dataset. Dimensionality of the input set is reduced from 35 variables to only 5. It can be noted that prediction of planar defects is less accurate than the prediction of other observables; however, a larger number of principal components could be selected to ensure the error in predictions is reduced.

### Table 4.10: Sensitivity indices for (a) $C_{11}$ and (b) $E_{\text{grain-boundary}}$ formation energy.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.00 %</td>
</tr>
<tr>
<td>$X_2$</td>
<td>11.52 %</td>
</tr>
<tr>
<td>$X_3$</td>
<td>10.96 %</td>
</tr>
<tr>
<td>$X_4$</td>
<td>10.97 %</td>
</tr>
<tr>
<td>$X_5$</td>
<td>10.95 %</td>
</tr>
<tr>
<td>$X_6$</td>
<td>10.98 %</td>
</tr>
<tr>
<td>$X_7$</td>
<td>10.93 %</td>
</tr>
<tr>
<td>$X_8$</td>
<td>11.19 %</td>
</tr>
<tr>
<td>$X_9$</td>
<td>11.01 %</td>
</tr>
</tbody>
</table>

(a) $C_{11}$ elastic constant

(b) $E_{\text{grain-boundary}}$ formation energy
Figure 4.11: Principal component analysis combined with RBF surrogate and error for (a) $C_{11}$, (b) $C_{12}$, elastic constants, (c) $E_{\text{vacancy}}$, (d) $E_{\text{interstitial}}$, (e) $E_{\text{grain-boundary}}$, and (f) $E_{\text{free-surface}}$ formation energies when using a variability of ±5% in the EAM potential for aluminium material system.
Chapter 5

Interatomic Potential from Surrogate Models

The purpose of this chapter is to carry out a feasibility study into the application of surrogate modeling techniques to speed-up a class of computationally expensive inverse problems in MD. First, the objectives of this chapter are outlined, followed by a complete description of the problem statement. Finally, the results of the study and the concluding remarks of this approach are presented.

5.1 Inverse Problem: Objectives

The studies presented in the previous chapters discussed the effects of introducing uncertainty in the interatomic potential parameters to analyze the sensitivity on the material properties of different material systems. Using the methodology discussed in Chapter 2, surrogate models were constructed to predict the quantities of interest based on a set of data arising from the parameters of the interatomic potentials and the results obtained from MD simulations. This can be viewed as a forward problem, i.e., given input parameters predict a set of output properties. However, surrogate models can also be used to solve the inverse problem where the interatomic potential parameters are to be estimated given a set of output quantities (observables) that are obtained via experiments or \textit{ab initio} calculations.

Reiterating the problem statement described in section 1.2, interatomic potentials are created via the application of optimization based fitting procedures using a set of experimental or \textit{ab initio} data. However, most of these procedures employ high weights for the parameters or properties of interest and do not account for any kind of experimental noise. For instance, in the case of graphene the fitting procedure was designed to deter-
mined the interatomic potential parameters according to the individual terms described in equation 3.1. As described in [66], the parameters of the REBO term are kept at their original values and are only modified due to the presence of the additional terms in order to model those observables with high weight. The parameters of the LJ term, on the other hand, were determined by using properties obtained from graphite material systems. Similarly, the parameters from the TORSION term were derived from the properties of butane. In a more general sense, fitting of interatomic potentials involves performing an optimization process to find the parameters that match the experimental observables to the calculated material properties from a prescribed functional form [20, 95, 96]. In the case of the metals EAM potential, the structure of the potential is defined by the functional form described in equation 4.3 but the fitting procedure is based on a set of cubic splines that offer a large number of parameters for an optimized fit [90]. In order to calibrate and validate the interatomic potential, MD simulations could be employed as proposed by [97] in order to account for atomic relaxations. However, the validation of the interatomic potentials can become computationally infeasible due to the large number of MD simulations required in order to optimize the potential [12].

Since the development of interatomic potentials is based on an optimization problem subjected to certain constraints, surrogate models can help to speed up the optimization process. Similarly, the use of surrogates offer the flexibility of accounting for non-smooth or noisy responses, and can further provide insights into the nature of the design space [98].

The goal of this chapter is to investigate the application of surrogates to efficiently solve optimization problems encountered in MD simulations. For this analysis, the objective is to recover the interatomic potential parameters from a given set of observables using the Gaussian process surrogate and minimizing the error function using a simulated annealing algorithm. To illustrate the proposed approach, graphene is selected as the material of choice; however, the approach studied here is general in scope and can be applied to other material systems.

5.2 Problem Statement

This section describes the formulation of the inverse problem and provides an overview of the methodology. Consider an interatomic potential model parametrized in terms of \( \mathbf{x} \in \mathbb{R}^d \), where \( d \) represents the dimension of the problem or in this case the number of parameters to recover. Since the goal is to recover the parameters of the interatomic potential, the estimation of \( \mathbf{x} \) referred to as \( \mathbf{x}^* \) is required such that the MD predictions
of a set of observables $O_{i}^{MD}(x)$ for $i = 1, \ldots, m$ agree with the experimental values of the observables $O_{i}^{exp}$ for $i = 1, \ldots, m$. The discrepancy or error function between the MD observables and the experimental values is defined as follows:

$$f(x) = \sum_{i=1}^{m} \frac{(O_{i}^{MD} - O_{i}^{exp})^2}{O_{i}^{exp}} \quad (5.1)$$

The above least square error metric is normalized using the observables from the experimental dataset to avoid any scale invariants between the observables. The inverse problem studied in the present work can be posed as a general bound constrained nonlinear programming problem as follows:

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x_l \leq x \leq x_u
\end{align*} \quad (5.2)$$

where $x_l$ and $x_u$ represent the lower and upper bounds constraints respectively. Since each evaluation of the objective function involves a full MD simulation, the optimization problem under consideration is computationally expensive.

### 5.2.1 Procedure

The following steps outline the general procedure followed to conduct this study and solve the inverse problem:

1. Carry out design of experiments (DOE) to generate a set of points in the parameter space. Carry out MD simulations at each point to create a training dataset that contains candidate parameter values along with the corresponding values of the objective function.

2. Construct and validate a Gaussian process surrogate model of the form $\hat{f}(x)$ that maps the input parameters to the objective function.

3. Optimize the surrogate model to estimate $x^*$.

4. Carry out MD simulation at $x^*$ to assess quality of parameter estimates. If convergence is not reached, append this new point to the training dataset and go back to Step 2.

In the present study, the pseudo-experimental values of observables were used so that the exact solution is known beforehand. The DOE dataset was taken from the previous study.
performed in Chapter 3. In addition, simulated annealing was chosen as the optimization algorithm to estimate the interatomic potential parameters of the pseudo-experimental dataset. Multiple initial guesses are used in the procedure in order to investigate if the problem is well-posed.

5.2.2 Simulated Annealing Algorithm

The simulated annealing algorithm is used to find the global optimum of a function in the presence of several local optima. It is based on the premise that slowly cooling crystalline structures after being heated up to high temperature, would allow the system to reach a minimum energy configuration [47]. The algorithm works by exploring the whole functional space in a direction that decreases a given cost function. By applying random changes on the evaluation point of this function or by using a probability obtained from the Boltzmann distribution it is possible to capture the optimum value. In general, the algorithm is described as follows:

<table>
<thead>
<tr>
<th>Algorithm 1: Simulated annealing algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data: ( x ) ← Initial guess</td>
</tr>
<tr>
<td>Result: ( x ) ← Optimized solution</td>
</tr>
<tr>
<td>( i ) ← 0;</td>
</tr>
<tr>
<td>while Global condition NOT satisfied do</td>
</tr>
<tr>
<td>( T ) ← CoolingSchedule(( i ));</td>
</tr>
<tr>
<td>( i ) ← ( i ) + 1;</td>
</tr>
<tr>
<td>while Local condition NOT satisfied do</td>
</tr>
<tr>
<td>( x^* ) ← RandomNeighbour(( x ));</td>
</tr>
<tr>
<td>( \Delta E ) ← ( F(x^*) - F(x) );</td>
</tr>
<tr>
<td>if ( \Delta E &lt; 0 ) then</td>
</tr>
<tr>
<td>( x ) ← ( x^* );</td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td>( r ) ← RandomVal[0,1];</td>
</tr>
<tr>
<td>if ( r &lt; \exp^{-\Delta E/kT} ) then</td>
</tr>
<tr>
<td>( x ) ← ( x^* );</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>end</td>
</tr>
</tbody>
</table>

The Boltzman probability distribution is given by \( P(\Delta E) = \exp^{-\frac{\Delta E}{kT}} \), and the cooling schedule is in charge of determining the temperature \( T \) which is the main control parameter of the process. As explained by [47], the initial temperature is usually chosen to be
high in order to keep almost all changes, but as temperature decreases only improved values are kept to reach the desired optimum.

5.3 Study: Recovering Graphene’s Potential

The initial study to determine the interatomic potential parameters was performed using four observables from the graphene dataset. The observables included are: the two elastic constants $C_{11}$, $C_{12}$, the cohesive energy, and the vacancy formation energy. The input parameters of the pseudo-experimental dataset are based on the functional form of the AIREBO potential, i.e., two parameters, $Q_{CC}$ and $\alpha_{CC}$ from the repulsive term and three parameters $\beta_{CC1}$, $\beta_{CC2}$, and $\beta_{CC3}$ from the attractive term of the REBO potential; four parameters arising from the LJ term: $r_{CC}^{LJ,min}$, $r_{CC}^{LJ,max}$, $\epsilon_{CC}$, and $\sigma_{CC}$; and the $\epsilon_{CCCC}$ parameter from the TORSION potential. The MD dataset consists of $N = 500$ samples and convergence is tested in the study by using different initial guesses to the optimization algorithm. Table 5.1 shows a comparison of the true interatomic potential parameters of the pseudo-experimental dataset and those obtained from solving the optimization problem. From the results shown, it can be noted that using four observables provides a good approximation to the true values of the pseudo-experimental parameters. After several iterations, the average error in the estimation of the input parameters is calculated, i.e., $||x - x^*|| = 6.7618e^{-4}$. From one set of the obtained input parameters, the four observables were reconstructed using MD simulations as well as the surrogate models already described in Chapter 3 to ensure the predicted parameters of the interatomic potential are able to provide a reasonable estimation of the quantities of

<table>
<thead>
<tr>
<th></th>
<th>True Value</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{CC}$</td>
<td>0.3139695</td>
<td>0.3105509</td>
<td>0.3158026</td>
<td>0.3122408</td>
</tr>
<tr>
<td>$\alpha_{CC}$</td>
<td>4.7057762</td>
<td>4.7023249</td>
<td>4.7004309</td>
<td>4.7100731</td>
</tr>
<tr>
<td>$\beta_{CC1}$</td>
<td>4.7085800</td>
<td>4.698978</td>
<td>4.6984736</td>
<td>4.7148985</td>
</tr>
<tr>
<td>$\beta_{CC2}$</td>
<td>1.4198439</td>
<td>1.4466303</td>
<td>1.4403432</td>
<td>1.4228983</td>
</tr>
<tr>
<td>$\beta_{CC3}$</td>
<td>1.3800957</td>
<td>1.3952075</td>
<td>1.3962312</td>
<td>1.3953366</td>
</tr>
<tr>
<td>$r_{CC}^{LJ,min}$</td>
<td>3.4051160</td>
<td>3.4322296</td>
<td>3.4062097</td>
<td>3.3728935</td>
</tr>
<tr>
<td>$r_{CC}^{LJ,max}$</td>
<td>3.8221139</td>
<td>3.8522425</td>
<td>3.8239187</td>
<td>3.7863924</td>
</tr>
<tr>
<td>$\epsilon_{CC}$</td>
<td>2.852716e^{-3}</td>
<td>2.816278e^{-3}</td>
<td>2.82700e^{-3}</td>
<td>2.86461e^{-3}</td>
</tr>
<tr>
<td>$\sigma_{CC}$</td>
<td>3.4051160</td>
<td>3.4326504</td>
<td>3.4055665</td>
<td>3.3732306</td>
</tr>
<tr>
<td>$\epsilon_{CCCC}$</td>
<td>0.3108159</td>
<td>0.3074496</td>
<td>0.3071090</td>
<td>0.3091850</td>
</tr>
</tbody>
</table>
interest as compared to those from the pseudo-experimental dataset. Results are shown in Table 5.2.

Table 5.2: Comparison between MD simulation and GPM results when using recovered input parameters

<table>
<thead>
<tr>
<th></th>
<th>$C_{11}$</th>
<th>$C_{12}$</th>
<th>$E_{\text{cohesive}}$</th>
<th>$E_{\text{vacancy}}$</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O^{\text{exp}}$</td>
<td>903.41 GPa</td>
<td>129.30 GPa</td>
<td>-6.55 eV</td>
<td>6.73 eV</td>
<td>-</td>
</tr>
<tr>
<td>$O^{\text{MD}}(x^*)$</td>
<td>912.05 GPa</td>
<td>126.53 GPa</td>
<td>-6.49 eV</td>
<td>6.79 eV</td>
<td>0.142199</td>
</tr>
<tr>
<td>$O^{\text{GPM}}(x^*)$</td>
<td>912.24 GPa</td>
<td>126.53 GPa</td>
<td>-6.50 eV</td>
<td>6.70 eV</td>
<td>0.145421</td>
</tr>
</tbody>
</table>

Simulation times were also captured as part of the comparison process. To perform this comparison, a small number of atoms was used for the simulations of the four observables and the time of the simulation was captured after completed. For a single observable prediction, the average MD simulation time was $t_{MD} = 1.2$ hr. In comparison, the surrogate model average times are $t_{\text{train}}^{\text{GPM}} = 4.45$ s and $t_{\text{predict}}^{\text{GPM}} = 0.28$ s for training and to compute single predictions respectively.

In order to provide a more exhaustive analysis, a similar study was performed using additional observables. For the case of five observables, the bond rotation formation energy is added; when six observables are present, the divacancy formation energy is added. Finally, when seven observables are used, the grain boundary formation energy is included in the analysis. Similar to the previous study, a few initial guesses were used to ensure the optimization problem is well-posed. Table 5.3 shows the results and the error in the predicted input parameters.

Table 5.3: Solution of optimization problem using five, six, and seven observables

<table>
<thead>
<tr>
<th></th>
<th>True Value</th>
<th>5 Observables</th>
<th>6 Observables</th>
<th>7 Observables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{\text{CC}}$</td>
<td>0.3130944</td>
<td>0.3108763</td>
<td>0.3105505</td>
<td>0.3107721</td>
</tr>
<tr>
<td>$\alpha_{\text{CC}}$</td>
<td>4.7729143</td>
<td>4.7919473</td>
<td>4.7929537</td>
<td>4.7930548</td>
</tr>
<tr>
<td>$\beta_{\text{CC}}$</td>
<td>4.7270098</td>
<td>4.7482427</td>
<td>4.7483139</td>
<td>4.7488654</td>
</tr>
<tr>
<td>$\beta_{\text{CC}_2}$</td>
<td>1.4194859</td>
<td>1.4210062</td>
<td>1.4192772</td>
<td>1.4262987</td>
</tr>
<tr>
<td>$\beta_{\text{CC}_3}$</td>
<td>1.3830306</td>
<td>1.3701750</td>
<td>1.3729268</td>
<td>1.3706190</td>
</tr>
<tr>
<td>$r_{\text{CC}}^{\text{LJmin}}$</td>
<td>3.380949</td>
<td>3.3708007</td>
<td>3.3906187</td>
<td>3.3779633</td>
</tr>
<tr>
<td>$r_{\text{CC}}^{\text{LJmax}}$</td>
<td>3.7949868</td>
<td>3.7800128</td>
<td>3.7978667</td>
<td>3.7942929</td>
</tr>
<tr>
<td>$\epsilon_{\text{CC}}$</td>
<td>2.83847e-3</td>
<td>2.85847e-3</td>
<td>2.85923e-3</td>
<td>2.86854e-3</td>
</tr>
<tr>
<td>$\sigma_{\text{CC}}$</td>
<td>3.380949</td>
<td>3.3787293</td>
<td>3.3869523</td>
<td>3.3826024</td>
</tr>
<tr>
<td>$\epsilon_{\text{CCC}}^{\text{T}}$</td>
<td>0.3062253</td>
<td>0.3097394</td>
<td>0.3099440</td>
<td>0.3067983</td>
</tr>
<tr>
<td>Error: $|x - x^*|$</td>
<td>-</td>
<td>3.8990e-4</td>
<td>3.6047e-4</td>
<td>3.5231e-4</td>
</tr>
</tbody>
</table>

It can be noted that the error in the prediction of the potential parameters is small. Comparing the individual parameters it is possible to conclude that regardless of the
initial guess, the system seems to provide a good approximation to the input variables as compared to those in the experimental dataset.

### 5.4 Study: Parameter Estimation with Noisy Observables

The objective of this study is to solve the optimization problem when experimental noise is present. That is, when the observables $y$ arising from experimental procedures are corrupted by additive measurement noise as follows:

$$ y = \mathcal{G}(x) + \eta $$  \hspace{1cm} (5.3)

where $\eta$ represent the noise in the measurement process $\mathcal{G}$. In general, the value of the noise parameter is not known, but some statistical properties can be inferred. In order to accomplish the proposed objective, the observables from the experimental dataset described in section 5.2 are subjected to small variations in their values prior to running the optimization algorithm. The aim of this exercise is to understand the level of noise that can be tolerated when recovering the input parameters of the experimental dataset.

To perform the analysis, noise is introduced to the seven observables of the pseudo-experimental dataset in variations of 2%, 5%, 10%, and 20% from the nominal value in order to study the sensitivity of the parameter estimates to noise. Figure 5.1 shows the histograms displaying the error percentage when noise is introduced to the experimental observables. The error is calculated using the L2-norm from the difference between the predicted input parameters and the experimental ones.

The histograms show the error spread when different variations of noise are present. For the case of 2% noise, it can be noted that there is a high probability to obtain accurate predictions, with a similar error percentage as described in the previous section where no noise is introduced. When noise is increased to 5%, the spread in the error is larger and the percentage of a higher error is also increased. For higher values of noise, the histogram shifts towards values of higher error values, meaning that there is a high probability of a higher error. In fact, the average error shifts from 0.0423% to 0.0612% for noise variations of 2% to 20% respectively; almost a 43% increase in error. Although the error is small and it is possible to conclude that the methodology is robust enough to capture the parameters of the interatomic potential, this analysis demonstrates that introducing noise to experimental observables increases the error in the predicted values of the interatomic potential parameters and should be something to consider when
Figure 5.1: Histograms displaying error percentage when noise is introduced to the experimental observables prior to the prediction of the input parameters. Dotted line represents the error mean of the distribution.

5.5 Concluding Remarks

The feasibility study in this chapter showed the application of surrogate models to calibrate the parameters of the interatomic potential of graphene from a set of pseudo-experimental observables. In practice, the observables can be obtained from a set of experiments or *ab initio* calculations and this methodology could be applied to obtain a set of interatomic potentials capable of describing those observables even when noise is present in experimental procedures. Although this work was performed with up to
seven observables, a larger set of observables could be used to estimate a larger set of potential parameters. By applying this methodology it was found that the predictions of the interatomic potential parameters is robust and leads to a good estimation of the observables as compared to those from the pseudo-experimental dataset. Similarly, it was shown that the surrogate models used in this work are able to make fast and accurate predictions of the observables.

The introduction to the inverse problem in this chapter aims to point future research in the right direction in terms of the creation of a more efficient interatomic potential capable of predicting the statistical distribution of the material properties. In a more extensive approach, the optimization problem should generate a probability distribution function (PDF) for the estimation of the input parameters $\mathbf{x}^*$ of the potential. To implement such methodology, a Bayesian framework can be used to obtain the posterior distribution of the parameters based on a priori knowledge obtained from a dataset arising from observations or experimental procedures [99]. In contrast to current methodologies where single values for the observables are used to calibrate the interatomic potential, the Bayesian approach can account for the ignored uncertainty that originates during experimental or computational measurements. Using Bayesian inference, the PDF of the parameters $\mathbf{x}^*$ is updated based on measurements or MD simulation.

One complication of this formulation is that the Bayesian inference cannot guarantee that the posterior distribution is known, even for large datasets for which the central limit theorem motivates a Gaussian posterior distribution. In fact, in most cases the predicted posterior distribution is an arbitrary distribution with an unknown normalization factor making sampling a very challenging task. To address this issue, Markov Chain Monte Carlo (MCMC) algorithms can be used, as they provide means of sampling from a probability measured only up to a normalized constant [12]. The difficulty in the implementation of MCMC techniques for the inverse problem is usually related to the complexity of the forward model. That is, the evaluation of the likelihood function to make the required inference. This step is computationally expensive as it requires multiple MD simulations. This motivates the use of surrogate models as an alternative to compute the likelihood function. Similarly, for the cases where only single point estimations are required, this approach is still valid as it is possible to compute the maximum a posteriori (MAP) estimate for the set of input parameters to which MD simulations or surrogate models can predict the estimation of the required observable.
Chapter 6

Conclusion and Future Work

This chapter summarizes the findings of applying surrogate models to characterize material properties when variability is introduced in the interatomic potentials used during molecular dynamics simulations. In this thesis, Chapters 3 and 4 use graphene and aluminium as material systems to investigate the sensitivity of their output properties when uncertainty is introduced in the system. Additionally, a methodology is proposed in Chapter 5 to solve an optimization problem using surrogates to determine the parameters of the interatomic potential given a set of observables. Furthermore, future research directions are proposed.

6.1 Conclusions

From the problem statement describe in Chapter 1, the following objectives were proposed:

1. Reproduce material topologies (with and without defects) to perform MD simulations in order to predict material properties.
2. Develop methodologies to propagate and quantify uncertainty in the parameters of interatomic potentials.
3. Formulate and validate surrogate models to evaluate the effects of uncertainty in interatomic potential parameters on materials properties predicted by MD simulations.
4. Utilize surrogate models to accelerate the solution of the inverse problem in order to determine the parameters of the interatomic potential based on observables.

These objectives were accomplished using surrogates such as the Bayesian Gaussian process models and radial basis functions with Gaussian kernels. The surrogates were then
applied to two different material systems with different interatomic potentials to study the sensitivity of the material properties to the interatomic potential parameters.

### 6.1.1 Summary of Methodology

The methodology chapter covered the methods and techniques used in this thesis. Starting with the theory of MD simulations, it was shown that MD relies on Newtonian physics to solve the many-body problem. Controls for temperature and pressure were also introduced to provide a complete description of how atomic simulations are performed.

Methods to propagate and quantify uncertainty were also presented. In particular, Monte Carlo, sparse grids, and Latin hypercube methods are introduced to sample the parameter space and estimate the statistical moments of the outputs of the MD simulations. The advantages of each of the techniques are discussed by presenting the relative error or the order of the method. For instance, the error for the Monte Carlo method is given by $\frac{\sigma}{\sqrt{N}}$ which is independent of the dimensionality of the problem, but very dependent on the number of samples. In comparison, the accuracy of sparse grids is dependent on the dimensionality of the problem, and also the number of points in the quadrature rule.

The chapter also introduces the concept of surrogate models as a method of approximating computationally expensive models. This work covers two basic surrogate models: The Gaussian process model combined with Bayesian statistics and the Radial Basis Functions with Gaussian kernels. The Bayesian statistics are used to estimate the shape parameters of the GPM, and to make inferences on the predictions of unseen points in the parameter space. RBFs are introduced as a comparison model to the GPM to show the underlying flexibility of GPM. Moreover, reduced order modelling using principal component analysis is presented to further enhance the application of surrogate models by reducing the dimensionality of the input space for cases where large number of inputs are used.

The mathematical formulation of how the surrogates are validated is also presented. Techniques such as the leave-one-out, leave-p-out, and k-fold cross validation are shown to discuss the effectiveness of the models. Similarly, the standardized cross validation residuals are presented as a metric to compute the confidence interval of the models. The plots showing the validation of the models is presented in Chapters 3 and 4 as they are application dependent. After ensuring that the models are valid, they are used to perform a sensitivity analysis where the main effects and interaction effects among variables can be computed.
6.1.2 Graphene

Graphene was selected as one of the material systems of choice to apply the proposed methodology. Initially, the graphene topology is discussed as well as the AIREBO interatomic potential used to describe the material when performing MD simulations. The topology for graphene is created by arranging atoms in a hexagonal ring and ensuring that periodic boundary conditions are met in all directions. The topology is also generated for graphene sheets containing defects such as vacancy, bond rotation, and grain boundaries. The MD topology of defective graphene shows similar configurations as those shown in literature. A complete examination on how the material properties are obtained is also presented in order to provide the reader with some insights on how MD simulations are performed.

Since the AIREBO potential includes a term based on the LJ potential to account for the intermolecular interactions, an initial study was perform to corroborate studies that can be found in the literature. From the analysis it was shown that a high variability (5%) does not amount for a large variation in the quantities of interest which is in agreement with similar studies in the literature [11, 12]. The subsequent study was performed using a black-box approach with ten parameters from the AIREBO potential as inputs and eight observables. Monte Carlo and sparse grids were used to sample the parameter space at different variabilities to showcase the effects on the output properties when uncertainty was introduced to the input parameters. The analysis showed that the observables are very sensitive to the input as the error in the prediction of the MD simulations were high and inconsistent. By sampling the full design space the robustness of the interatomic potential was shown to decrease at high values of variability.

Upon showing that the surrogate models are optimal for this application, a sensitivity analysis was presented to show how each of the variables in the input parameter affects each of the observables. It was shown that mainly two parameters are responsible for the large spread found in the quantities of interest. This effects are also quantified in terms of the Sobol indices showing that the responsible parameters affect the observables approximately by 40% more than all other parameters.

6.1.3 Aluminium

A similar process to that described in the previous section is followed for the application to aluminium material systems. The topology for this material is created using the LAMMPS software commands. Since the metallic structure is well known, it is possible to also include point, line, and planar defects such as vacancies, stacking faults, and grain
boundaries to name a few. Compared to graphene, the EAM interatomic potentials used for metallic systems are usually obtained in a tabulated form instead of its functional form. In this approach, the input parameters to the black-box are obtained from the coefficients of the fitting process. The dimensionality of the problem is then obtained by the number of coefficients that cause the less amount of error during the fitting.

The use of surrogate models for this application is motivated by the fact that the parameter space is very sensitive to uncertainty. By performing Monte Carlo sampling it was noted that observables that were used to construct the EAM interatomic potential show a large amount of error of the order of approximately 43% for the worst case scenario when only a variability of 5% is applied to the input terms. For this material system, only the GPM surrogate was used since it provided the best resolution. Furthermore, the model was able to accurately predict any unseen points from the design space. Quantification of the sensitivity on the observables was performed using main effects and Sobol indices which, due to the nature of the problem showed a similar effect for most of the input variables.

The EAM potential has a very high dimensional parameter space since the potential is given by 10000 points for each of the functions it represents. In order to see the effects of two of the functions of the EAM potential, a reduced order model is employed to reduce a 35 input set to a set of only 5 variables with low error in the predictions. The implementation of this model is validated using 100 points in the training set and 400 points for testing.

\subsection{Inverse Problem}

Numerical studies were also conducted to investigate the feasibility of accelerating the solution of inverse problems in MD using surrogate models. The goal of this study is to recover the input parameters of some “experimental” dataset by minimizing the error function using the Gaussian process model. The cost function was based on the least square error between the pseudo-experimental observables and the observables from the MD simulations. Several studies were performed by increasing the number of observables used in the error function to verify if the problem is ill-posed. At the same time, different initial guesses was selected for each of the trials to verify that the optimizer algorithm, based on simulated annealing techniques, could evolve to find the global optimum of the problem instead of any possible local minima. Using the set of observables, it was possible to recover the input terms of the interatomic potential from the experimental dataset within a small error. Moreover, MD simulations were performed to verify that the
predicted parameters would provide an accurate estimation of the quantities of interest as compared to those observables from the experimental dataset.

The effect of noise on the parameter estimates were also studied numerically. These studies suggest that the interatomic potential parameters can be estimated even when noise is present. Noise was introduced in variations of 2%, 5%, 10%, and 20% to the nominal values of the pseudo-experimental observables. It was noted that the mean error in the prediction of the input parameters of the pseudo-experimental dataset increased by 43% when noise variations shifted from 2% to 20%.

6.2 Future Work

This section outlines some avenues for future research in the topic of uncertainty quantification of MD simulations.

6.2.1 MD Simulations

Even though there seems to be a growing interest in combining surrogate models with applications that require MD simulations, some work is still necessary to develop techniques that could accurately describe the properties of material systems when uncertainty is present. Current techniques to create interatomic potentials, for instance, are based on fitting methodologies that make use of noisy data from experimental procedures.

Future work in molecular dynamics can be done to understand the effects of temperature or ensembles on the sensitivity of material properties. Many material properties are highly dependent on this variable and therefore, it becomes important to further analyze how temperature affects the propagation of uncertainty during MD simulations. Similarly, combining MD with ab initio simulations can be further improved by employing statistical techniques to deal with different sources of uncertainty. Although this work was applied to graphene and aluminium systems, several other material systems could be tested with the same methodology proposed to understand how sensitive their properties are when uncertainty is present. Furthermore, different interatomic potentials under uncertainty could be compared to investigate how each potential affects the sensitivity of the material properties and therefore, estimate which potentials provide a more robust and reliable prediction of the quantities of interest.
6.2.2 Surrogate Models

This work presented the use of Gaussian process model and radial basis functions as the surrogate models of choice. However, a more comprehensive comparison should be performed between several other surrogate models in order to capture computational complexity and performance. In this work, a comparison between RBFs with Gaussian kernels and GPMs was made, but in many cases it is more convenient to use RBFs with other kernels to ensure a better performance not only in the prediction of unseen points but also in terms of computational resources. Throughout this thesis, one of the assumptions in the formulation of the uncertainty quantification was the idea that the probability distribution function (PDF) of the input parameters of the interatomic potential is known. However, future studies should employ more realistic models for the PDFs of the input parameters prior to introducing uncertainty and observing the sensitivity of the material properties.

The motivation for the employment of surrogate models to solve the inverse problem was presented in this work. Using a Bayesian approach combined with advanced sampling techniques such as the Markov Chain Monte Carlo, it is possible to obtain a prediction of the PDF for the parameters of the interatomic potential based on observations arising from experiments or ab initio calculations [12]. The use of a technique of this sort, is computationally expensive as it requires a large number of MD simulations to provide a good minimization of the error between observed and predicted values. However, by applying surrogate models as described in Chapter 5, it is possible to reduce computational time and produce a robust estimation of the interatomic potential parameters. In addition, when single point estimates are required this approach can make use of the maximum a posteriori (MAP) estimate to obtain a set of input parameters that would produce the required set of observables [99].
Bibliography


