Aprendizaje Automático de Potenciales Intermoleculares para Diseñar Materiales para Refrigeración Ecológica

Tesis presentada a la **Universidad Autónoma de Baja California**
como requerimiento para obtener el grado de:
Maestro en ciencias (MC)

Por
Miguel de Jesus Chavez Hernandez
Matricula: 1246569

Supervisor
Dr. Mauricio Alonso Sanchez Herrera
Facultad de Ciencias Químicas e Ingeniería
Universidad Autónoma de Baja California

Septiembre, 2024

Queen Mary University of London

Machine Learning Intermolecular Potentials to Design Materials for Green Refrigeration

A thesis submitted to the Queen Mary University of London
for the degree of Master of Philosophy (MPhil)

By
Miguel de Jesus Chavez Hernandez
Student ID: 221185081

Supervisor
Professor Anthony E. Phillips
School of Physical and Chemical Sciences
Queen Mary University of London

September, 2024

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS QUÍMICAS E INGENIERÍA

Folio No.369
Tijuana, B.C., a 13 de noviembre, 2024

C. Miguel de Jesús Chávez Hernández
Pasante de: Maestría en Ciencias
Presente

POR LA REALIZACIÓN PLENA DEL SER

El tema de trabajo y/o tesis para su examen profesional, en la

Opción TESIS .

Es propuesto, por el C. Dr. Mauricio Alonso Sánchez Herrera .

Quien será responsable de la calidad del trabajo que usted presente, referido al tema "Aprendizaje Automático de Potenciales Intermoleculares para Diseñar Materiales para Refrigeración Ecológica" .

El cual deberá usted desarrollar, de acuerdo con el siguiente orden:

I. INTRODUCCIÓN
II. ANTECEDENTES
III. METODOLOGÍA
IV. RESULTADOS Y DISCUSIONES
V. CONCLUSIONES Y TRABAJO FUTURO.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
POR LA REALIZACIÓN PLENA DEL SER
FACULTAD DE CIENCIAS QUÍMICAS E INGENIERÍA CAMPUS TIJUANA

M.C. Roberto Alejandro Reyes Martínez
Director

Dra. Ana Alejandra Ramírez Rodríguez
Subdirectora

Dr. Mauricio Alonso Sánchez Herrera
Director De Tesis

Declaración de Colaboración Institucional

Esta tesis ha sido realizada en el marco de un programa de doble titulación de maestría en investigación, fruto de la colaboración académica entre la Universidad Autónoma de Baja California (UABC) y Queen Mary University of London (QMUL). Este esfuerzo colaborativo permitió desarrollar una investigación integral que forma parte tanto del programa Research Master's (ResM) de QMUL como del programa Maestría en Ciencias e Ingeniería de UABC.

El trabajo de investigación presentado en esta tesis fue llevado a cabo durante un período de dos años. En el primer año, completado en UABC, se realizó un estudio profundo de las herramientas y de la teoría fundamental necesarias para el desarrollo de este proyecto de investigación, titulado "Aprendizaje Automático de Potenciales Intermoleculares para Diseñar Materiales para Refrigeración Ecológica". La principal área de estudio abordada fue la química computacional, la cual se complementó con conocimientos en física, química cuántica y ciencias de la computación. Esta etapa inicial del proyecto, enfocada en el establecimiento de los fundamentos teóricos y metodológicos, estuvo bajo la supervisión del Dr. Mauricio Alonso Sánchez Herrera (UABC), el Dr. Anthony Phillips (QMUL), quienes brindaron orientación clave para definir el marco y los objetivos de la investigación.

Durante el segundo año, el proyecto fue acogido por QMUL, donde se profundizó en la fase experimental y se implementaron modelos avanzados de aprendizaje automático que requirieron el uso de tecnología de vanguardia y recursos especializados. Esta etapa incluyó reunir datos de simulaciones computacionales diversas en interacciones químicas, optimización de las variables a utilizar en estas simulaciones, comparación con datos recolectados en el acelerador de partículas ISIS de neutrones y muones así como la implementación de diversos algoritmos y arquitecturas de aprendizaje automático. Bajo la dirección y supervisión del Dr. Anthony Phillips y el Dr. Alston J. Misquitta en QMUL como asesor y co-asesor respectivamente, se lograron avances significativos que permitieron abordar el problema de manera integral obteniendo simulaciones meramente computacionales fieles a los resultados obtenidos por el acelerador de partículas ISIS, fortaleciendo así los resultados y aplicabilidad de la investigación en la utilización de aprendizaje automático.

La presente tesis cumple con los requisitos académicos y científicos de ambas instituciones y permite una evaluacion integral en ambas universidades. Reflejando los altos estándares de rigor que estas instituciones exigen para la obtención de la doble titulación. Esta colaboración internacional es un esfuerzo conjunto que busca fomentar la excelencia académica y proporcionar una experiencia investigativa enriquecida, aprovechando la sinergia entre los recursos y conocimientos especializados de ambas universidades.

Este documento, que constituye el esfuerzo conjunto de ambas instituciones y el fruto de la dedicación académica vertida este tiempo, quedará registrado en los archivos académicos de UABC y QMUL, reflejando un compromiso con la innovación científica y la cooperación internacional en la formación de investigadores.

Contents

Abstract

This project addresses the challenge of modeling interatomic interactions in the barocaloric material quinuclidinium hexafluorophosphate (Quin-PF6) through the development of a custom forcefield (FF). To enhance the accuracy and adaptability of predictions, a neural network was employed to learn from the forcefield, with the ultimate goal of determining its potential to replace classical forcefields in molecular dynamics simulation software. The forcefield was trained using trajectory data directly from molecular dynamics simulations as the primary source of information, involving the fitting of high-quality data obtained from *ab initio* calculations. A comprehensive comparison between simulation results and real-world behavior was conducted to validate the model.

The study demonstrated the effectiveness of the developed forcefield in capturing complex intermolecular interactions and energy variations under pressure changes. Key objectives included performing high-precision *ab initio* calculations, employing advanced fitting techniques for the Buckingham potential, and generating a robust training dataset from molecular trajectory data. The results revealed a strong correlation between the energies calculated using SAPT2 and those predicted by the optimized forcefield.

While the optimization of intermolecular interactions was successful, particularly for PF6-PF6 and QUIN-QUIN interactions, challenges remained in accurately modeling the QUIN-PF6 system due to its complexity. Future work will focus on refining molecular interaction representations using advanced descriptors, improving the dataset, and exploring hybrid models that combine machine learning with traditional physical approaches. The findings of this research pave the way for deeper insights into the barocaloric properties of Quin-PF6 and the broader implications for ionic materials.

Acknowledgments

Dedicated to my beloved mother, whose unwavering support and love have been my guiding light.

I would like to express my heartfelt gratitude to CONAHCYT,
Autonomous University of Baja California,
and Queen Mary University of London for their generous funding.

I am also deeply thankful to my friends and family as well as everyone who has helped me along this journey. Your encouragement and belief in me made this possible.

Statement

I, Miguel, hereby confirm that the research presented in this thesis is my own work. Where collaborative efforts or support from others have been involved, appropriate acknowledgement is given below, and my specific contributions are clearly indicated. Any previously published material has also been acknowledged. I affirm that I have taken due care to ensure the originality of this work, and to the best of my knowledge, it does not violate any UK laws, infringe upon third-party copyrights or Intellectual Property Rights, nor contain confidential information.

I understand that the University reserves the right to use plagiarism detection software to review the electronic version of this thesis.

I also confirm that this thesis has not been submitted for a degree at this or any other institution. The copyright of this thesis belongs to the author, and no part of it may be quoted or used without my prior written consent.

Signature: Miguel de Jesus Chavez Hernandez
Date: 26/09/23

1 Introduction

1.1 Motivations

The growing global demand for refrigeration systems, driven by increasing populations and the intensifying effects of climate change, highlights the urgent need for sustainable, eco-friendly alternatives. Traditional vapor-compression refrigeration systems, while ubiquitous and efficient, rely heavily on refrigerants with high global warming potential and consume vast amounts of energy, predominantly generated from fossil fuels. This not only exacerbates the greenhouse effect but also contributes to the rapid depletion of natural resources. As environmental regulations tighten and the search for greener technologies intensifies, it has become imperative to explore innovative materials and methods that can reduce the environmental footprint of cooling systems[1] [2].

Barocaloric materials, which exhibit entropy changes in response to pressure variations, offer a promising solution. These materials can potentially eliminate the need for harmful refrigerants and significantly reduce energy consumption, addressing both ecological and energy efficiency concerns. Furthermore, recent research on ionic plastic crystals (IPCs), specifically quinuclidinium hexafluorophosphate (Quin-PF6), has demonstrated colossal barocaloric effects, positioning them as strong candidates for next-generation solid-state refrigeration technologies[3, 4].

The working principle of barocaloric systems lies in the application or release of pressure, which induces significant entropy changes that can be harnessed for efficient cooling. These entropy changes are driven by reversible phase transitions between disordered and ordered states within the material. Unlike traditional refrigerants, which rely on chemical transitions, barocaloric materials undergo physical changes, allowing for a more sustainable and energy-efficient cooling process. The barocaloric effect is akin to the magnetocaloric and electrocaloric effects but is more advantageous for practical applications due to the mechanical nature of pressure, which is easier to control and apply in a variety of systems. This makes barocaloric materials particularly appealing for next-generation refrigeration technologies, especially as their performance improves with recent advances in material synthesis and characterization[5].

The motivation for this research lies in the potential to revolutionize the refrigeration industry by developing a high-precision, neural network-based forcefield (FF) tailored to Quin-PF6. By leveraging advanced computational methods such as molecular dynamics (MD) simulations, ab initio calculations, and machine learning techniques, this project aims to model the complex interatomic interactions within barocaloric materials more accurately than ever before. Ultimately, the goal is to produce a forcefield that not only enhances the understanding of these materials but also sets the stage for practical applications in eco-friendly cooling technologies. This effort could have far-reaching implications for energy efficiency, environmental protection, and the development of future cooling devices.

1.2 Aims

In this project, the challenge of modeling interatomic interactions in the barocaloric material quinuclidinium hexafluorophosphate (Quin-PF6) is addressed by developing a custom forcefield (FF). To improve the accuracy and adaptability of the predictions, a neural network will be employed to learn from the forcefield with the ultimate goal of verify if it can replace the clasical forcefield in molecular dynamics simulation softwares in the future. To achieve this the forcefield will be trianed, using trajectory data directly from molecular dynamics simulations as the primary source of information. This will involve fitting a forcefield to high quality data obtained from ab initio calculations in order to be exceptionally precise. To verify it, comparations between the results generated in the simulations and the real world behaviour of the system are going to be performed. Through this methodology, complex intermolecular interactions and energy variations under pressure changes can be captured in greater detail, improving the predictive capabilities of the Neural network based forcefield and enhancing the exactness of the predictions.

1.3 Objectives

The main objectives of the project are the following:

1. Perform high-precision ab initio calculations on the Quin-PF6 system to obtain the necessary parameters that will enable the fitting and refinement of the forcefield using high-quality theoretical data.
2. Employ advanced fitting techniques to precisely calibrate the parameters of the Buckingham potential.
3. Collect molecular trajectory data obtained from molecular dynamics simulations to generate an accurate training dataset that serves as the foundation for developing the neural network-based forcefield.
4. Perform a comprehensive comparison between simulation results and real-world experimental data to assess the degree of alignment, ensuring the model accurately replicates physical behaviors.
5. Develop a set of transformations from molecular dynamics simulation data to generate representative descriptors that capture the key characteristics of intermolecular interactions in the Quin-PF6 system.
6. Use the obtained descriptors to train a neural network capable of accurately predicting the energy behavior of the Quin-PF6 system in molecular simulations.
7. Experiment with various neural network architectures, including different layer configurations, activation functions, and optimizers, to identify the structure that yields the highest predictive accuracy for the Quin-PF6 system.

2 Background

2.1 Materials and properties

2.1.1 Introduction to the barocaloric effect and plastic crystals

In the search for more sustainable and eco-friendly alternatives to combat climate change, solid-state refrigeration systems have emerged as a promising option [6]. These systems, which do not rely on greenhouse gases or traditional mechanical compressors, offer a greener and more efficient approach to refrigeration [7]. The absence of harmful refrigerants, such as hydrofluorocarbons (HFCs), significantly reduces their environmental impact. Additionally, their reliance on solid-state materials allows for a more direct conversion of energy into cooling effects increasing the efficiency and saving energy. In these systems, various physical phenomena are key to inducing temperature and entropy changes through the application of external stimuli. Among them, caloric effects, such as the magnetocaloric, electrocaloric, elastocaloric, and barocaloric effects, play a crucial role[8].

- The **magnetocaloric effect** occurs when a material experiences an entropy change in response to an applied magnetic field, causing a temperature change when subjected to adiabatic conditions.
- The **electrocaloric effect** is analogous, where an electric field induces entropy changes, commonly found in certain ferroelectric materials.
- The **elastocaloric effect** involves mechanical stress (like tension or compression) that induces a temperature change through the entropy changes associated with elastic deformations in the material.
- The **ionocaloric effect** occurs when the introduction of ions into the surroundings of a material lowers its melting point via an electrochemical field, triggering a solid-liquid phase transition. This endothermic process absorbs energy, resulting in cooling as the material melts[9].
- Lastly, the **barocaloric effect (BCE)** is defined by an entropy change in response to hydrostatic pressure, which is the focus of the system studied in this work.

The easy implementation of refrigeration systems based on these materials supports the idea that barocaloric materials are promising candidates for solid-state refrigeration. The pressure-induced temperature change in these materials can be integrated into various technologies with relative simplicity. One of the most promising systems for generating this effect is **plastic crystals**, also known as **orientationally disordered crystals**, particularly those that undergo **orientational order-disorder solid-solid phase transitions** [5, 10]. These materials combine an ordered crystalline structure with the rotational freedom of molecules or ions, enabling them to exhibit unique barocaloric behavior under applied pressure.

Among plastic crystals, a subcategory of particular interest is **ionic plastic crystals**. Unlike molecular plastic crystals, in which neutral molecules rotate freely and are primarily influenced by weak interactions such as hydrogen bonds and van der Waals forces, ionic plastic crystals are composed of **cations and anions** that still retain high rotational or vibrational mobility [11]. This combination of order, due to their ionic nature, and dynamic disorder, due to rotational freedom, bestows them with thermodynamic properties that are highly beneficial for solid-state refrigeration. Notably, these systems have demonstrated significant entropy changes, which are critical for efficient cooling [8, 12].

2.1.2 Barocaloric Effect and its Thermodynamic Foundations

The barocaloric effect (BCE) is defined as the change in a material's entropy in response to an applied hydrostatic pressure. In barocaloric materials (BCMs), pressure can be applied adiabatically, leading to a temperature change, which is a key characteristic that positions these materials as promising candidates for solid-state refrigeration technologies [13]. In all caloric materials, large entropy changes are typically associated with significant alterations in the material's degrees of freedom. For the BCE, the coupling between these degrees of freedom and an external stimulus, pressure in this case, results in entropy changes.

The thermodynamic description of this effect is based on the first law of thermodynamics:

$$du = TdS - pdV \tag{2.1}$$

where du is the change in internal energy, T is the temperature, dS is the change in entropy, p is the applied pressure, and dV is the change in volume. The entropy change associated with the BCE depends largely on how the volume of the material responds to the applied pressure. When the material undergoes a volumetric change, the lattice degrees of freedom are modified, and this results in an increase or decrease in entropy.

Constant Volume Condition

However, if the volume is held constant during the pressure application, the behavior of the system differs significantly. Under constant volume conditions ($dV = 0$), the thermodynamic expression simplifies to:

$$du = TdS \tag{2.2}$$

In this case, no work is done by or on the system due to volume changes, as $pdV = 0$. Therefore, any entropy change would no longer be mediated by the lattice volume but could still occur due to other degrees of freedom, such as electronic or vibrational excitations [14]. However, these contributions are typically smaller compared to the entropy changes associated with lattice rearrangements.

In barocaloric materials, the lack of volumetric changes would significantly diminish the extent of the entropy change.(Δs_T), as the BCE relies heavily on the coupling between pressure and lattice volume. This highlights the importance of compressibility in BCMs; materials with higher compressibility will exhibit larger entropy changes under hydrostatic pressure, making them more suitable for barocaloric applications. On the other hand, when the volume is constrained, the effectiveness of pressure-induced entropy changes, and consequently temperature changes, is diminished.

Maxwell Relations and Specific Heat

From the thermodynamic relations, we can derive the following Maxwell relation, which directly connects entropy and volume changes:

$$\left(\frac{\partial S}{\partial p}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_p \tag{2.3}$$

This equation illustrates the relationship between the change in entropy concerning pressure at a constant temperature and the change in volume concerning temperature at a constant pressure. This connection is essential for comprehending how entropy fluctuates when pressure is applied to or removed from a barocaloric material.

By integrating this relation over a pressure range, we can express the total change in entropy as:

$$\Delta s_T = \int_{p_0}^{p_f} \left(\frac{\partial S}{\partial p}\right)_T dp = -\int_{p_0}^{p_f} \left(\frac{\partial V}{\partial T}\right)_p dp \tag{2.4}$$

Additionally, the specific heat at constant pressure, c_p, is commonly defined as the derivative of enthalpy with respect to temperature at constant pressure:

$$c_p = \left(\frac{\partial H}{\partial T}\right)_p \tag{2.5}$$

From the relation $dH = TdS + VdP$, this leads to the following result for the specific heat in terms of entropy:

$$c_p = T\left(\frac{\partial s}{\partial T}\right)_p \tag{2.6}$$

This specific heat expression links the temperature dependence of entropy to heat capacity, a critical factor when studying how pressure-induced entropy changes translate into thermal effects.

Adiabatic Temperature Change in BCMs

When a BCM is subjected to a pressure change under adiabatic conditions, the system experiences a temperature change as a direct result of the entropy variation. This temperature change (ΔT_{ad}) can be computed as:

$$\Delta T_{ad} = \int_{p_0}^{p_f} \frac{T}{c_p}\left(\frac{\partial s}{\partial p}\right)_T dp \tag{2.7}$$

The ability of a BCM to undergo temperature changes due to pressure variations is what makes these materials attractive for refrigeration applications, where an inverse BCE may also be observed (i.e., cooling upon pressure release) [14, 15].

2.1.3 Current Barocaloric Materials and the Relevance of Ionic Plastic Crystals

Barocaloric materials (BCMs) are a diverse group of substances exhibiting significant entropy changes under applied pressure, a feature that makes them promising candidates for solid-state refrigeration technologies. Various categories of materials show barocaloric effects (BCE), including shape memory alloys, spin-crossover complexes, plastic crystals, hybrid organic-inorganic materials, ferroelastics, ferroelectrics, natural rubber, polymers, and n-alkanes [16, 17, 18, 19, 20, 21]. Among these, plastic crystals have attracted considerable attention due to their distinctive structural properties, particularly

their orientational disorder. Orientationally disordered (OD) crystals, commonly referred to as plastic crystals, are characterized by molecules arranged in a regular lattice but with substantial orientational disorder [10].

This unique combination of structural order and rotational freedom creates an intermediate phase between liquids and traditional crystalline solids. Plastic crystals can undergo solid-solid phase transitions, often accompanied by significant entropy changes. The barocaloric effects in these materials come from phase transitions between ordered and disordered states, where the change in entropy mainly comes from the disorder caused by the rotational freedom of the molecules. This disorder is triggered or adjusted by applying pressure, resulting in significant barocaloric effects.

The shape of the molecules in plastic crystals is important for their barocaloric behavior. Molecules with spherical characteristics, which facilitate rotational freedom, contribute to enhanced entropy changes as demonstrated by studies involving molecules such as methane and adamantane [10].

A significant group of plastic crystals that has recently attracted interest is ionic plastic crystals (IPCs). IPCs are composed of cations and anions, leading to much stronger Coulombic interactions than those observed in neutral molecules found in traditional plastic crystals. These enhanced interactions not only stabilize the crystal structure but also greatly influence the material's thermal and mechanical behavior, making them crucial for applications in energy storage and refrigeration. Despite these stronger ionic interactions, IPCs retain a high degree of rotational or vibrational freedom, which is particularly advantageous for inducing large entropy changes under pressure [22]. This combination of strong ionic interactions with the ease of molecular rotation improves the barocaloric effect. Studies have shown that the rotational mobility of ions in these materials can result in colossal entropy changes, enhancing their potential for cooling applications [12], with performance similar to that of the common R-134A refrigerant. Furthermore, the combination of ionic conduction and barocaloric effects opens the door to multifunctional applications, such as in energy storage or thermal management systems [22].

2.1.4 Promising ionic plastic crystals: Quin-PF6

One particularly promising example of an IPC is quinuclidinium hexafluorophosphate (Quin-PF_6). Quinuclidinium is a globular organic cation related to DABCO (1,4-Diazabicyclo[2.2.2]octane), but with one nitrogen atom replaced by carbon, resulting in a molecule with a threefold rotational axis. In this transformation, the molecule loses a mirror plane and three twofold rotational axes, leading to a reduction in symmetry. This descent can be formally described as a reduction from the D_{3h} point group to the C_{3v} point group, or equivalently, from $\bar{6}m2$ to 3m. This structure promotes a high degree of orientational freedom, which, when combined with the ionic nature of the molecule, produces significant configurational entropy.

The phase transition in Quin-PF6 from an orthorhombic ordered phase to a high-temperature cubic plastic phase demonstrates a substantial entropy change, making this material highly relevant for barocaloric applications. Our research group has collected data indicating that Quin-PF6 undergoes an order-disorder transition at around 298 K under ambient pressure, a temperature range suitable for room-temperature cooling applications. Additionally, the ionic conduction properties of Quin-PF6, typical of organic ionic plastic crystals (OIPCs), suggest its potential as a multifunctional material with applications extending beyond refrigeration into areas such as solid-state electrolytes for batteries and fuel cells [23]. These combined properties of high barocaloric strength, reversible entropy change, and ionic conduction underscore Quin-PF6's significance as a compelling candidate for advanced research and practical applications being one of the main motivations of this thesis.

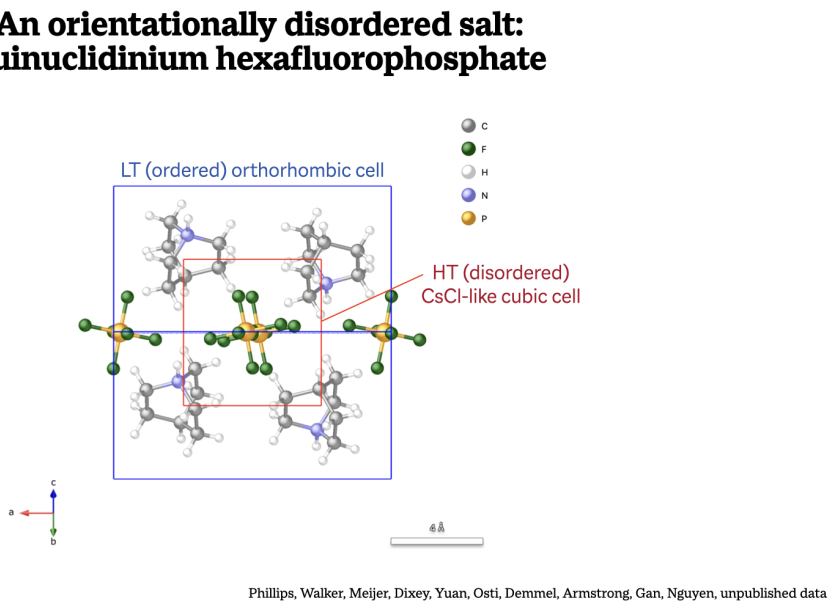


Figure 2.1: Quinuclidinium hexafluorophosphate

2.2 Molecular Dynamics

2.2.1 Introduction to molecular dynamics

We carry out computer simulations to understand how the structure and the microscopic interactions allow some molecules to have specific properties. Normally we do this to have a complement to the experiments in order to broad the information of the system and learn something new from it. One of the most used and well established methods is molecular dynamics. Molecular dynamics (MD) simulations is a computational technique that has become one of the most useful and reliable tools for modelling materials at the atomistic level, serving as a bridge between microscopic length and time scales and the macroscopic world with highly accurate predictions of the system evolution. MD dynamics works in a very simple way; it predicts the evolution of a system by solving Newton's second law of motion for each individual particle. By applying force fields, which are mathematical models describing the potential energy of interactions between particles, Molecular Dynamics simulations calculate the forces acting on each particle and integrate their equations of motion over discrete time steps. This method provides a detailed trajectory of the system's behaviour over time, allowing for the exploration of complex interactions and dynamics within the system. It allows for the investigation of specific phenomena such as phase transitions and mechanical responses, and facilitates the refinement of theoretical models to better understand and optimise these materials for multiple applications [24].

2.2.2 Forcefields and intermolecular potentials

A force field (FF) is a mathematical expression that establishes the dependency of the system's energy on the coordinates of its particles. Basically, we construct this expression using a finite set of parameters to match the physics of the system. The values of these parameters and the form of the expression depend on exact calculations, usually based on quantum chemistry principles, and an understanding of the nature of the system[25] Although they can also be fitted using experimental data.

The Force field replaces the true potential with a simplified model that is as close as possible to the original value. This is why the values used to adjust the parameters need to be as accurate as possible[26].

Parameterization methods are crucial in defining the FF. Parameters can be determined through various approaches, including fitting to experimental data, high-level quantum calculations, or a combination of both. These methods aim to refine the force field so that it accurately represents the system's behavior under different conditions.

If the FF has a good set of parameters that accurately represents the system, it is possible to analyze the properties and behavior of the system in many configurations. This potential energy is then used to calculate the forces acting on each atom and allows the MD program to simulate the system's evolution in defined time steps[27]. For accurate molecular dynamics simulations, it is common to use well-established force fields (FF) that are specifically fitted for certain types of molecules or families of materials. However, if the system under study is not well-defined yet, these "general" force fields may not perform adequately, leading to potential errors in calculations. Examples of established force fields include CHARMM36m[28], AMOEBA[29], OPLS, OPLS-AA [30], AMBER[31], and GROMOS[32] among others. Each of these is specialized for different types of systems.

A typical form of the force field equation can be expressed as follows:

$$\begin{aligned} U = &\sum_{\text{bonds}} \frac{1}{2} k_b \left(r - r_0\right)^2 \quad + \sum_{\text{angles}} \frac{1}{2} k_a \left(\theta - \theta_0\right)^2 + \sum_{\text{torsions}} \frac{V_n}{2} \left[1 + \cos(n\phi - \delta)\right] \\ &+ \sum_{\text{improper}} V_{imp} \quad + \sum_{\text{LJ}} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right) + \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}} \end{aligned} \tag{2.8}$$

Where the FF terms represent the following:

- **Bond Term:** This term represents the harmonic potential associated with the stretching or compression of bonds between atoms. It is given by $\frac{1}{2}k_b(r - r_0)^2$, where k_b is the bond force constant, r is the bond length, and r_0 is the equilibrium bond length.
- **Angle Term:** This term accounts for the bending of bonds away from their equilibrium angles. It is described by $\frac{1}{2}k_a(\theta - \theta_0)^2$, where k_a is the force constant for angle bending, θ is the actual angle, and θ_0 is the equilibrium angle.
- **Torsion Term:** This term describes the energy associated with the rotational distortion of the dihedral angle between two planes of atoms. It is given by $\frac{V_n}{2}[1 + \cos(n\phi - \delta)]$, where V_n is the torsional force constant, n is the multiplicity of the torsion, ϕ is the dihedral angle, and δ is the phase shift.
- **Improper Term:** This term represents the energy associated with deviations from planarity or non-standard torsions in molecular structures. It is denoted by V_{imp}, representing the force constant for such interactions.
- **Lennard-Jones (LJ) Term:** This term models the van der Waals interactions between non-bonded atoms. It is given by $4\epsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right)$, where ϵ_{ij} is the depth of the potential well, σ_{ij} is the finite distance at which the potential is zero, and r_{ij} is the distance between atoms i and j.
- **Electrostatic Term:** This term represents the Coulombic interactions between charged particles. It is expressed as $\frac{q_i q_j}{r_{ij}}$, where q_i and q_j are the charges of atoms i and j, and r_{ij} is the distance between these charges.

When experimenting with a new system, it is often necessary to deviate from classical descriptions and modify the mathematical representation of one or more terms. This adaptation is crucial for capturing the specific characteristics of the system, allowing for a more accurate and realistic analysis and modeling[25].

It is important to note that while the Lennard-Jones potential is a common choice for modeling interatomic interactions due to its simplicity and reasonable accuracy in many cases, depending on the specific system, other interatomic potentials may offer better representations. These can include, but are not limited to [33, 34, 35]:

- **Morse Potential:**

$$V(r) = D_e \left[\exp\left(-2\alpha(r - r_0)\right) - 2\exp\left(-\alpha(r - r_0)\right)\right] \tag{2.9}$$

  This potential describes the interaction between two atoms with a potential well depth D_e, where α controls the width of the potential well and r_0 is the equilibrium bond distance.

- **Buckingham Potential:**

$$E_{jk} = B\exp\left(-Cr_{jk}\right) - Ar_{jk}^{-6} \tag{2.10}$$

  This potential combines an exponential term for short-range repulsion with a r^{-6} term for long-range attraction, where A and ρ control the repulsive term, and C is the coefficient for the attractive term.

- **Stillinger-Weber Potential:**

$$V(r_{ij}) = \phi_{ij}(r_{ij}) + \sum_{k\neq j} \psi_{ijk}(r_{ij}, r_{ik}, \theta_{ijk}) \tag{2.11}$$

  This potential includes two-body and three-body terms to model interactions, particularly in covalent systems like silicon, where $\phi_{ij}(r_{ij})$ represents the pairwise interaction and $\psi_{ijk}(r_{ij}, r_{ik}, \theta_{ijk})$ accounts for the three-body interactions.

Choosing the right term for modeling interatomic interactions is crucial, as the forces governing molecular interactions directly determine the physical properties of a material. An incorrect choice of potential can lead to an inaccurate physical representation of the system. Intermolecular interactions, including Van der Waals, induction, dispersion, and electrostatic forces, play a fundamental role in force fields, as they can contribute significantly to the system's total energy. According to Stone (2013), these forces not only influence the structural stability of molecules but also affect macroscopic properties such as cohesion and phase behaviour [36].

2.2.3 Equations of motion

In molecular dynamics simulations, the primary goal is to describe the time evolution of a system of particles under the influence of forces acting between them. The equations of motion, derived from Newton's second law, allow us to calculate the trajectories of particles by integrating the forces present in the system.

The mathematical foundation of molecular dynamics simulations relies on determining the total energy of the system, which depends on the positions and velocities of the particles. Specifically, the potential energy of the system is used to calculate the forces acting on each particle. These forces, in turn, dictate how the particles interact with each other and influence their motion over time[37].

Newton's Equations

The equation of motion for a particle in a system is based on Newton's second law, expressed as:

$$\mathbf{F}_i = m_i \mathbf{a}_i \tag{2.12}$$

Where:

- $\mathbf{F}_i$ is the force acting on particle i,
- m_i is the mass of particle i,
- $\mathbf{a}_i = \frac{d^2\mathbf{r}_i}{dt^2}$ is the acceleration of particle i.

Newton's equation can be rewritten as a second-order differential equation in terms of the particle position $\mathbf{r}_i$:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{2.13}$$

Potential Energy and Forces

The force on each particle can be obtained from the derivative of the total potential energy $V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ of the system with respect to the position of particle i:

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} V \tag{2.14}$$

Where:

- V is the potential energy, which may include pairwise interactions, many-body interactions, and other terms such as electrostatics or dispersion.
- $\nabla_{\mathbf{r}_i}$ denotes the gradient with respect to the position of particle i.

The potential energy term is essential, as it describes how the forces depend on the relative positions of the particles, and these forces drive the system's dynamics.

Numerical Integration

To simulate the motion of the system, Newton's equations are solved numerically. The most common method for integrating the equations of motion is the Verlet algorithm. The Verlet algorithm is based on a Taylor series expansion for the particle positions:

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \Delta t^2 \mathbf{a}_i(t) \tag{2.15}$$

Where:

- $\mathbf{r}_i(t)$ is the position of particle i at time t,
- Δt is the time step,
- $\mathbf{a}_i(t)$ is the acceleration of particle i at time t.

This algorithm updates the positions of the particles at each time step based on the acceleration, which in turn is derived from the computed forces.

Energy Conservation

An important aspect of molecular dynamics simulations is the conservation of the system's total energy. The total energy E_{total} is the sum of the kinetic energy K and the potential energy V:

$$E_{\text{total}} = K + V \tag{2.16}$$

$$K = \sum_{i=1}^{N} \frac{1}{2} m_i \mathbf{v}_i^2 \tag{2.17}$$

Where:

- K is the kinetic energy,
- m_i is the mass of particle i,
- $\mathbf{v}_i$ is the velocity of particle i.

The balance between kinetic and potential energy is crucial to ensuring that the simulation remains physically accurate.

2.2.4 Simulation parameters

In molecular dynamics simulations, selecting appropriate simulation parameters is crucial for achieving accurate and efficient results. The time step (Δt) determines the frequency with which particle positions and velocities are updated. A smaller time step increases accuracy but also computational cost, with typical values ranging from femtoseconds (fs) to picoseconds (ps) depending on the system and forces involved. Temperature and pressure are key parameters that influence the system's thermodynamic properties and are often controlled using thermostats and barostats to maintain desired conditions throughout the simulation. The integration algorithm used affects the stability and accuracy of the simulation; common methods include the Verlet and Velocity Verlet algorithms, each offering different trade-offs in precision and computational efficiency. The cutoff radius defines the distance beyond which interactions between particles are ignored to reduce computational complexity. It must be chosen carefully to balance accuracy and efficiency. Lastly, the simulation length must be long enough to capture the dynamics of the system, influenced by the timescales of the processes being studied and the statistical convergence of the results [37, 25].

2.2.5 System specifications and ensembles

When performing molecular dynamics simulations, the most important aspect is to accurately represent the system in order to obtain reliable results. Various parameters in molecular dynamics simulations can vary depending on the system's characteristics, but not all of these necessarily model the system itself. Examples include simulation parameters that we have already mentioned. However, some specifications indeed directly affect how accurately the simulation mimics its real-world counterpart. These values are generally linked to the system's initial setup rather than the simulation parameters, as they describe the atoms, molecules, or proteins present in the system under study[38].

Describing the system is typically standardized depending on the software used. For example, when defining the geometry, atomic connections, protein chains, or other structural details of a molecule or system, specific file formats are employed. These file formats follow a standardized layout and language in text files with unique extensions, allowing the software to identify the type of file being

used. Examples include PDB (Protein Data Bank), XYZ, GRO (GROMACS), PSF (Protein Structure File), CIF (Crystallographic Information File), and MOL2, among others. All these formats share the common goal of describing the atoms in the system, their connections (all but CIF), types, and any labels related to chains or subgroups. The choice of format depends on the system type, the software best suited for the problem, personal preferences, and the level of detail required to describe the system [39, 40, 41].

Once the system is described in terms of atoms, connections, and subgroups, boundary conditions must be specified to emulate the environment in which the system exists. This may involve defining whether the system is surrounded by water, is an isolated molecule, involves a protein in contact with a cell, or represents a crystal that needs to be simulated as an infinite system with periodic boundaries. Regardless of the specific case, defining these boundary conditions is always necessary. Depending on the software, this can be done using additional specification files, graphical interfaces, or programming through the software's API[42].

The most crucial part of system creation and specification in molecular simulations is defining the forces and interactions between molecules and atoms. This involves choosing the force field, the mathematical representation of the forces in the system (potentials), and the parameter values within this force field. Users can often adjust the magnitude and influence of these forces in the final calculations. Well-established programs offer extensive features and flexibility for specifying these details, with some programs being more adaptable than others[36, 39].

Once the system is fully specified, and the forces and parameters are appropriately defined, attention must be given to how the software controls the conditions of the system. This is where *ensembles* come into play. Ensembles are theoretical frameworks that define the conditions under which the system evolves, such as maintaining constant temperature, pressure, energy, or particle number. The choice of ensemble is particularly important, as it influences the simulation outcomes and aligns with the desired experimental conditions[38].

Although ensembles are primarily concerned with the decision of maintaining specific conditions, they are enforced through mathematical formulas, algorithms, and physical laws. Therefore, depending on the ensemble chosen, different methods are applied to achieve the desired control within the simulation. Below is a brief overview of the most common ensembles and the formulas and algorithms typically used to impose them [42]:

- **NVE Ensemble (Microcanonical):** Maintains the number of particles (N), volume (V), and energy (E) constant.
  - *Mathematical Representation:* The equations of motion (Newton's laws) are integrated directly to conserve the total energy.
  - *Common Algorithm:* Velocity Verlet or Leapfrog integrators are used to update particle positions and velocities, ensuring energy conservation.
- **NVT Ensemble (Canonical):** Maintains the number of particles (N), volume (V), and temperature (T) constant.
  - *Mathematical Representation:* Temperature is controlled using an external heat bath.
  - *Common Algorithms:* Langevin dynamics and the Nosé-Hoover thermostat are commonly used to regulate the system's temperature, introducing damping and stochastic forces to maintain thermal equilibrium.

- **NPT Ensemble (Isothermal-Isobaric):** Maintains the number of particles (N), pressure (P), and temperature (T) constant.
  - *Mathematical Representation:* Both temperature and pressure are controlled to allow the system's volume to fluctuate.
  - *Common Algorithms:* The Nosé-Hoover thermostat (for temperature control) and the Parrinello-Rahman or Berendsen barostat (for pressure control) are often used.
- **μVT Ensemble (Grand Canonical):** Maintains the chemical potential (μ), volume (V), and temperature (T) constant.
  - *Mathematical Representation:* Particles can be exchanged with an external reservoir, altering the particle number to achieve equilibrium at a given chemical potential.
  - *Common Algorithms:* Monte Carlo moves are used to insert or delete particles in the system to maintain the chemical potential.

Each ensemble imposes a distinct set of constraints on the simulation, tailored to the expected outcomes of the experiment and the system's characteristics. These constraints guide the time evolution of the simulation to accurately reflect specific real-world conditions, thereby ensuring reliable results.

Some of the most commonly used software for molecular dynamics (MD) simulations are GROMACS, an open-source MD simulation package known for its performance and efficiency, particularly in biomolecular systems; AMBER, a suite of programs with a focus on biomolecules, providing various tools for force field development and analysis; CHARMM, widely used for MD simulations of biomolecular systems with extensive support for force fields and simulations; LAMMPS, a flexible and scalable tool suitable for a wide range of materials and scientific applications; and OpenMM, a high performance MD simulation library designed to be highly customizable and optimized for modern computing hardware [39, 42, 43, 44, 45] .

2.3 Quantum Chemistry

2.3.1 Introduction to quantum chemistry

Quantum chemistry is a fundamental discipline within the chemical sciences that applies the principles of quantum mechanics to explore and predict the properties and behaviors of molecular systems. By focusing on the interactions between electrons and nuclei, quantum chemistry enables a detailed understanding of electronic structures, molecular geometries, reaction pathways, and spectroscopic properties. This is achieved by solving the Schrödinger equation, which provides a quantum mechanical description of molecular systems that goes beyond the approximations of classical models. Modern quantum chemical methods, including ab initio approaches such as Hartree-Fock (HF), Density Functional Theory (DFT), and post-Hartree-Fock techniques like Symmetry-Adapted Perturbation Theory (SAPT), have become essential tools in both theoretical studies and practical applications. These methods offer valuable insights into chemical processes, facilitating advancements in fields ranging from materials science to biochemistry, and thereby enhancing the predictive power and scope of chemical research [46, 47, 48].

2.3.2 Ab initio methods in quantum chemistry

Ab initio methods are a fundamental approach in quantum chemistry, aiming to calculate the properties of molecular systems based solely on the principles of quantum mechanics, without relying on experimental or empirical data. The term "ab initio," which translates from Latin as "from the beginning," highlights the fact that these methods depend only on basic physical constants and the underlying mathematical framework to obtain accurate and relevant results[49].

In ab initio methods, the electronic Schrödinger equation is solved for a given system to describe the electronic structure and predict various properties, such as molecular geometries, vibrational frequencies, and reaction energies. Unlike empirical methods, which use experimentally derived parameters, ab initio approaches employ theoretical models to provide a detailed understanding of molecular interactions.

The most basic ab initio method is the Hartree-Fock (HF) theory, where the wavefunction of the system is approximated as a single determinant of molecular orbitals. More sophisticated methods, such as post-Hartree-Fock approaches, introduce electron correlation effects to improve accuracy. Although ab initio calculations can be computationally intensive, they are crucial for exploring the properties of molecules where experimental data is scarce or for validating empirical models[49].

2.3.3 Schrödinger equation

The Schrödinger equation is a key equation in quantum chemistry that helps us to understand how quantum systems behave. Although finding exact solutions can be really tough for complex systems, it is the starting point for many approximate methods used to study these systems[50].

Time-Dependent Schrödinger Equation

The time-dependent Schrödinger equation is given by:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = \hat{H}\Psi(\mathbf{r},t) \tag{2.18}$$

where:

- $\Psi(\mathbf{r},t)$ is the wave function of the system, which contains all the information about the system's state.
- $\hat{H}$ is the Hamiltonian operator, representing the total energy of the system.
- $\hbar$ is the reduced Planck constant.
- i is the imaginary unit.

Time-Independent Schrödinger Equation

For systems where the Hamiltonian does not explicitly depend on time, the time-independent Schrödinger equation is used:

$$\hat{H}\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{2.19}$$

where:

- $\hat{H}$ is the Hamiltonian operator, representing the total energy of the system.

- $\psi(\mathbf{r})$ is the spatial part of the wave function.
- E is the energy eigenvalue associated with the state described by $\psi(\mathbf{r})$.

Wave Functions for more than one electron

We can utilize the wave function to determine the likelihood of finding a particle at a specific location. For a single spinless particle in one dimension, the function $|\Psi(\mathbf{r})|^2$ is particularly useful. Provides the probability density for the location of the particle in position r, which is crucial to predict where the particle is most likely to be found. This probabilistic interpretation of the wave function allows us to make quantitative predictions about the behavior of particles, $|\Psi(\mathbf{r})|^2$ can also be expressed as:

$$|\Psi(\mathbf{r})|^2 = \Psi^*(\mathbf{r})\Psi(\mathbf{r}) = \rho(\mathbf{r}) \tag{2.20}$$

The electron density $\rho(\mathbf{r})$ quantifies the likelihood of finding an electron near a particular location in space. It is expressed mathematically as:

$$\rho(\mathbf{r}) = \langle\Psi|\hat{\rho}|\Psi\rangle \tag{2.21}$$

where the density operator $\hat{\rho}$ is defined as:

$$\hat{\rho}(\mathbf{r}) = \sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{r_i}) \tag{2.22}$$

In this context, N represents the total number of electrons, while δ denotes the Dirac delta function, which reflects the contribution of each individual electron.

When the Schrödinger equation is extended to systems with more than one electron, the complexity of the problem increases significantly. For many-electron systems, the general acceptable wave functions can be made by forming the linear superposition of all permutations of the electron variables, with the appropriate sign [49]:

$$\psi\left(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\right) = C \sum_P (-1)^P \hat{P} \psi_a\left(\mathbf{x}_1\right) \psi_b\left(\mathbf{x}_2\right) \ldots \psi_n\left(\mathbf{x}_N\right) \tag{2.23}$$

where $\hat{P}$ is an operator that permutes the variables $\mathbf{x}_1 \ldots \mathbf{x}_N$ and we add all the permutations. The sign $(-1)^P$ is positive if P consists of an even number of interchanges, and is negative if the number of interchanges is odd. The constant C is only needed to normalize the terms.

Operator for kinetic and potential energy

When dealing with quantum mechanics it is important to know that all the observable quantities are represented by operators. The operator for kinetic energy can be denoted by $\hat{T}$ for kinetic energy of a single electron with the following mathematical expression:

$$\hat{T} = -\frac{\hbar^2}{2m}\nabla^2 \tag{2.24}$$

Expanding this term to many electrons we can represent the total kinetic energy in re folloing way:

$$\hat{T} = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 \tag{2.25}$$

Where ∇_i^2 operates only on the position of electron i.

Electrons are attracted by the electrostatic force field of the nuclei. This can be represented in the following equation for the potential energy of an electron under this field. For a system of N electrons with positions $\mathbf{r}_i$, the total potential energy due to this phenomenon can be described as follows:

$$\hat{V} = \sum_{i=1}^{N} v\left(\mathbf{r}_i\right) \tag{2.26}$$

Where $\mathbf{r}$ is the position of the particle, $v(\mathbf{r})$ is the field, and $\hat{V}$ is the potential energy.

Electrons not only interact with the nuclei but also with each other as we can predict by Coulomb's law, so we can expect the potential energy between two electrons at positions $\mathbf{r}_1$ and $\mathbf{r}_2$ due to their mutual electrostatic interaction in a system of N electrons, represented by the operator $\hat{U}$ and expressed in the following way:

$$\hat{U} = \frac{1}{2} \sum_{i \neq j} \frac{e^2}{4\pi\epsilon_0 \left|\mathbf{r}_i - \mathbf{r}_j\right|} \tag{2.27}$$

The operator representing the total energy is the Hamiltonian $\hat{H}$, which is the sum of $\hat{T}$, $\hat{V}$, and $\hat{U}$:

$$\hat{H} = \hat{T} + \hat{V} + \hat{U}. \tag{2.28}$$

In a system with N electrons, the time-independent Schrödinger equation is expressed as:

$$\hat{H}\Psi_n\left(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\right) = E_n \Psi_n\left(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\right) \tag{2.29}$$

In simple terms, the Schrödinger equation predicts all possible states of a quantum system based on the system's Hamiltonian, which includes kinetic and potential energy terms. The wave function $\Psi(\mathbf{r}, t)$ describes the probability distribution of a particle's position and other properties, and by solving the equation, we obtain the system's allowed energy levels and corresponding wave functions. These solutions provide insights into the system's behaviour, such as the likelihood of finding particles in specific regions, which is crucial in solid-state physics for understanding the electronic structure of materials. While the equation forms the basis for many computational methods in quantum chemistry and materials science, solving it exactly can be challenging for systems with multiple interacting particles.

2.3.4 Hartree-Fock Method: the beginning

It is challenging to address the computational problems in quantum chemistry related to the Schrödinger equation due to the complexity and time required to solve it, making an exact solution virtually impossible for most systems. The Hartree-Fock approximation was the first practically useful method developed to tackle this issue.

The Hartree-Fock (HF) approximation is a method used to approximate the wave function of a many-electron system. It simplifies the problem of interacting electrons by assuming that each electron moves independently in an average field created by all other electrons. The wave function is approximated as a Slater determinant, which ensures that the wave function is antisymmetric with respect to the exchange of any two electrons. This approximation significantly reduces the computational complexity compared to solving the full Schrödinger equation[51, 52].

HF method is based on the variational principle, which states that the energy calculated from any

trial wave function is an upper bound to the true ground state energy of the system. This principle allows for the optimization of the trial wave function to find the best possible approximation to the ground state. Specifically, the goal is to minimize the total energy of the system with respect to the wave function parameters. Mathematically, the energy of the system is given by:

$$E = \sum_{i=1}^{N} \langle \phi_i | \hat{T} + \hat{V}_{\text{eff}} | \phi_i \rangle$$

where $\hat{T}$ represents the kinetic energy operator, which accounts for the motion of the electrons, and $\hat{V}_{\text{eff}}$ is the effective potential that includes the interactions between electrons and the external potential. The effective potential is expressed as:

$$\hat{V}_{\text{eff}} = \hat{V}_{\text{ext}} + \sum_{j=1}^{N} \left[\int \frac{|\phi_j(\mathbf{r})|^2}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r}' - \int \frac{\phi_j^*(\mathbf{r}) \phi_j(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r}' \right]$$

Here, $\hat{V}_{\text{ext}}$ is the external potential, which represents the interaction of electrons with an external field (such as the nuclei). The first term inside the sum is the Coulomb repulsion term, representing the classical electrostatic repulsion between the electron density of orbital ϕ_j and itself. The second term is the exchange term, which corrects for the Pauli exclusion principle by accounting for the exchange interactions between electrons in different orbitals.

The Hartree-Fock equations are solved iteratively, leading to a self-consistent field (SCF) solution where the orbitals and the energy converge. Despite its computational efficiency, the HF approximation does not account for electron correlation beyond the average field approximation, which is a limitation for accurately predicting certain properties. Advanced methods, such as post-Hartree-Fock approaches, aim to address these limitations by incorporating electron correlation effects.

We can say that basically The Hartree-Fock method calculates the total electronic wavefunction of a system by considering the exchange correlation among electrons arising from their antisymmetry property. This introduces an exchange term to correct the Hartree method. However, it still does not incorporate dynamic electron correlations. The single-particle Slater determinant wavefunction ignores Coulomb correlations, namely the spatial interactions due to electron repulsion. Post-Hartree-Fock techniques, like Configuration Interaction (CI), address these electron correlations more effectively. Additionally, a fundamentally different approach, Density Functional Theory, shifts focus from the wavefunction to electron density.

2.3.5 Density functional theory DFT

Density Functional Theory (DFT) is a powerful approach in quantum chemistry for studying many-electron systems. Unlike methods that solve the full Schrödinger equation, DFT focuses on the electron density $\rho(\mathbf{r})$ rather than the wave function $\Psi(\mathbf{r})$. This approach significantly simplifies calculations by reducing the problem to one of density instead of many-body wave functions.

This method for calculating electron density and energy relies on two key theorems proposed by Hohenberg and Kohn in 1964. The first theorem states that the external potential V_{ext} and the total energy of the ground state are unique functionals of the electron density. This means that the electron density $\rho(\mathbf{r})$ uniquely determines the external potential and the energy in the ground state of the system[53].

The second theorem asserts that the ground state energy can be obtained variationally. Specifically,

it states that the ground-state energy is the minimum value of the total energy functional when evaluated with the correct electron density. In other words, the density that minimizes the total energy functional $E[\rho]$ is the true density of the ground state. This variational approach ensures that the computed energy is an upper bound to the true ground state energy, and the true ground state is achieved when the density is optimized such that this energy is minimized[53].

Electron Density

In DFT, the electron density $\rho(\mathbf{r})$ is calculated from the Kohn-Sham orbitals $\psi_i(\mathbf{r})$, which are solutions to the Kohn-Sham equations. The density is given by:

$$\rho(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2 \tag{2.30}$$

where $\psi_i(\mathbf{r})$ are the Kohn-Sham orbitals, which represent an effective single-particle approximation to the many-body problem.

Kohn-Sham Equations

The Kohn-Sham equations transform the many-body problem into a set of single-particle Schrödinger-like equations. The equations are:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{eff}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}) \tag{2.31}$$

where $V_{eff}(\mathbf{r})$ is the effective potential, defined as:

$$V_{eff}(\mathbf{r}) = V_{ext}(\mathbf{r}) + V_{Hartree}(\mathbf{r}) + V_{xc}(\mathbf{r}) \tag{2.32}$$

where: $V_{ext}(\mathbf{r})$ is the external potential due to the nuclei, $V_{xc}(\mathbf{r})$ is the exchange-correlation potential, accounting for electron-electron interactions beyond the classical Coulomb interaction.

The Hartree potential $V_{Hartree}(\mathbf{r})$ is computed using:

$$V_{Hartree}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}' \tag{2.33}$$

This integral represents the Coulomb interaction of an electron with the electron density of the entire system. The integral is becomes a double integral when solving numerically due to the need to calculate the interaction between all pairs of points in the density distribution.

Basis Set Representation

In practical DFT calculations, the Kohn-Sham orbitals $\psi_i(\mathbf{r})$ are approximated using a basis set. The basis set consists of a set of functions $\phi_k(\mathbf{r})$ that are used to expand the orbitals. Specifically, the Kohn-Sham orbitals are represented as:

$$\psi_i(\mathbf{r}) = \sum_k c_{ik}\phi_k(\mathbf{r}) \tag{2.34}$$

where $\phi_k(\mathbf{r})$ are the basis functions and c_{ik} are the expansion coefficients for the i-th orbital. This expansion allows the complex Kohn-Sham equations to be solved in a practical computational manner.

Iterative Procedure

DFT calculations are performed iteratively to achieve self-consistency:

1. Assume an initial density distribution $\rho_{\text{in}}(\mathbf{r})$. This is often chosen to be a superposition of atomic electron densities.

2. Compute the Hartree potential $V_{\text{Har}}(\mathbf{r})$ and the exchange-correlation potential $V_{\text{xc}}(\mathbf{r})$, and hence the effective Kohn-Sham potential $V_{KS}(\mathbf{r})$:

$$V_{KS}(\mathbf{r}) = V_{ext}(\mathbf{r}) + V_{\text{Har}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r}) \tag{2.35}$$

3. Solve the Kohn-Sham equations to obtain the Kohn-Sham orbitals $\psi_n(\mathbf{r})$ and the Kohn-Sham orbital energies ϵ_n.

4. Fill up the lowest $\frac{1}{2}N$ states and compute the output density $\rho_{\text{out}}(\mathbf{r})$:

$$\rho_{\text{out}}(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2 \tag{2.36}$$

5. Compare $\rho_{\text{out}}(\mathbf{r})$ with $\rho_{\text{in}}(\mathbf{r})$. If they do not match within a specified tolerance, update $\rho_{\text{in}}(\mathbf{r})$ and repeat the process.

6. Finally, we used the self-consistent ground state density to compute the ground state energy.

This iterative procedure is essential for achieving a self-consistent solution in Density Functional Theory (DFT). By continuously refining the electron density and Kohn-Sham potentials, this approach ensures alignment between the calculated effective potential and the density from the Kohn-Sham equations. This not only improves the precision of the electron density, but also ensures that the Kohn-Sham orbitals and their energies accurately represent the electronic properties of the system[49].

2.3.6 Symmetry adapted perturbation theory SAPT

Symmetry-Adapted Perturbation Theory (SAPT) is an advanced quantum chemistry method utilized to explore the essential interactions between molecules. Distinct from other methods that necessitate separate monomer and dimer energy calculations, SAPT directly dissects the total interaction energy into physically interpretable components. These components—electrostatic, exchange, induction, and dispersion terms—offer deeper insights into the myriad factors influencing molecular interactions.

Within SAPT, "levels of theory" denote various approximations and truncations applied to the perturbation series to maintain computational manageability. These levels can be seen as analogous to basis sets in Density Functional Theory (DFT). In DFT, utilizing a larger, more complex basis set typically leads to more precision but also demands increased computational resources and time for convergence. Similarly, in SAPT, higher levels of theory mean more accurate calculations but also greater computational demands.

For instance, **SAPT0** is the most rudimentary level of theory, employing first-order approximations for exchange terms and excluding higher-order terms such as triple excitations. This level is appropriate for obtaining preliminary estimates of molecular interactions without significant computational expense. Conversely, **SAPT2** includes second-order corrections, enhancing the accuracy of interaction energy calculations, especially in systems where dispersion interactions are significant. **SAPT2+** adds further corrections refining the estimation of exchange and polarization terms, providing greater precision.

Lastly, **SAPT2+(3)** incorporates third-order approximations in specific terms, offering even more precision.[54]

It is critical to understand that advancing each theory level in SAPT not only boosts calculation accuracy but also raises computational requirements, paralleling the effect of extensive basis sets in DFT—yielding finer electronic distribution details at the cost of increased computational effort and time. Choosing the proper SAPT theory level involves balancing desired accuracy against available computational resources. This choice, akin to DFT, should ensure acceptable precision without excessive computational costs, particularly for large or complex molecular systems. [55]

The main formula for SAPT2 is:

$$E_{\text{SAPT2}} = E_{\text{elst}}^{(1)} + E_{\text{exch}}^{(1)} + E_{\text{ind}}^{(2)} + E_{\text{exch-ind}}^{(2)} + E_{\text{disp}}^{(2)} + E_{\text{exch-disp}}^{(2)} \tag{2.37}$$

where:

- $E_{\text{elst}}^{(1)}$ is the first-order electrostatic energy,
- $E_{\text{exch}}^{(1)}$ is the first-order exchange-repulsion energy,
- $E_{\text{ind}}^{(2)}$ is the second-order induction energy,
- $E_{\text{exch-ind}}^{(2)}$ is the second-order exchange-induction energy,
- $E_{\text{disp}}^{(2)}$ is the second-order dispersion energy,
- $E_{\text{exch-disp}}^{(2)}$ is the second-order exchange-dispersion energy.

2.4 Machine Learning

2.4.1 Introduction to Deep Learning

To talk about deep learning, it is pertinent first to discuss machine learning (ML). ML is a branch of artificial intelligence focused on enabling computers to learn and enhance their performance from extensive datasets without explicit programming. This field revolves around developing algorithms to detect patterns in data and construct models for particular tasks, facilitating precise predictions and intelligent actions [56].

Historically, machine learning has evolved from simple statistical methods to more complex algorithms capable of handling larger and more diverse datasets. Techniques such as linear regression, decision trees, and support vector machines have been widely used due to their interpretability and relatively low computational cost [57]. These methods work well for tasks where the relationship between input and output can be explicitly defined and where the data size is manageable [58].

Deep learning, a subset of machine learning, represents a significant advancement in this field. It leverages neural networks with many layers (hence "deep") to automatically learn hierarchical features from large amounts of data. This capability enables deep learning to excel in tasks such as image and speech recognition, natural language processing, and complex pattern recognition, such as predicting material and molecular properties [59]. Unlike traditional methods, deep learning models learn feature representations directly from raw data, often outperforming simpler models. However, they typically require more computational resources and longer training times.

Additionally, deep learning models are trained using optimization algorithms like gradient descent and its variants, which help minimize the error in predictions by adjusting model parameters. Regularization techniques, such as dropout and normalization, are also employed to prevent overfitting and improve generalization [60].

2.4.2 Data sources and preparation

When engaging in machine learning, the most critical factor, in my opinion, is the quality and reliability of the data. Despite referring to models as "intelligent," the reality is that they are essentially mathematical expressions operating in a multidimensional space. Thus, ensuring data quality is paramount to avoid complications during the learning process.

In today's data-driven era, acquiring data has become more sophisticated and ubiquitous. Companies often gather data directly from our devices through cookies, and analyze patterns from our social media interactions. This data collection is not just about convenience; it enables companies to refine their algorithms and tailor services based on user behavior[61, 62].

In fields such as physics and chemistry, data acquisition is even more complex. Instruments like particle accelerators, chromatography equipment, and medical devices generate vast amounts of data. Even seemingly simple sensors measuring humidity and temperature contribute to this data deluge. The variety and volume of data are immense, and extraction methods can range from direct measurements to more indirect means[59].

To manage this flood of data, scientists and engineers have developed sophisticated "data pipelines." A data pipeline is essentially a sequence of processes designed to extract information from raw sources, load it into a server, and store it for further analysis. These pipelines are crucial for transforming data into a usable format, and visualizing these processes through diagrams can help understand their complexity and efficiency[61, 63].

While data is abundant, ensuring its quality is a significant challenge. Addressing issues such as missing data points, erroneous measurements, and the necessity for consistent data processing is crucial. Questions like "Are there any missing data points?" "Should some data points be removed due to measurement errors?" "Is uniform processing necessary for all data?" and "Can scripts or models assist in data cleaning and management?" must be answered before proceeding with machine learning tasks.

These challenges highlight the need for expertise in data science and engineering. Properly preparing data sources and establishing efficient data flows are essential steps before diving into machine learning. The effectiveness of any model heavily relies on the foundation laid by a well-organized and high-quality dataset.

To analyze and manage data flows effectively, we can utilize a variety of tools and technologies. For scripting and automation, Python is widely used due to its powerful libraries like Pandas and NumPy, which facilitate data manipulation and analysis. SQL databases are essential for structured data storage and querying, while tools like Terraform help manage infrastructure as code, enabling scalable and reproducible data architectures. For data storage, options range from simple disk storage solutions to more sophisticated cloud storage services such as AWS S3, Google Cloud Storage, and Azure Blob Storage. Additionally, programs that generate data, such as simulation software and sensors, play a crucial role in data acquisition. Leveraging these technologies allows for efficient handling of data pipelines and ensures that data is readily available for analysis and modeling[64].

Under expert knowledge, maintaining a controlled data flow is always a good option. However, the nature of experiments and the variability in data acquisition times and operations will dictate the technologies used. While more complex technologies can be a powerful tool, they are not always necessary and simpler solutions may suffice depending on the situation (personal experience).

At this point, it is also important to highlight that data can be obtained from high-precision calculations, such as ab initio methods. These calculations offer highly accurate predictions and are valuable for various scientific and engineering applications. However, it is crucial to ensure that the programs and setups used for these calculations are sufficiently robust to exploit the data effectively. This techniques were employed in this project and will be detailed further in the methodology section to mention the specific approach used. By using advanced computational methods and validating their accuracy, we can ensure that the data we rely on is of the highest quality and suitability for our analyses[65, 66].

2.4.3 Artificial Neural Networks (ANNs)

In scientific research, the capability to forecast the behavior of complex systems is essential across multiple disciplines like physics and chemistry. Historically, models in these areas have predominantly used linear approximations and equilibrium assumptions, which often limit their capacity to accurately represent complex dynamics. Nonetheless, recent studies have revealed patterns and correlations within these systems that imply a degree of predictability. For instance, in physics, phenomena such as long-range correlations and clustering behaviors are observed in systems ranging from fluid dynamics to condensed matter. Similarly, in chemistry, molecular interaction patterns and reaction kinetics show structured behaviors that can be utilized for predictive modeling[67, 68].

Traditional scientific models, including those based on linear dynamics or equilibrium thermodynamics, have provided significant insights into system behaviors. However, these models generally fall short when dealing with non-linear interactions and dynamic complexities present in real-world systems. Examples include phase transitions, non-equilibrium states, and chaotic behaviors, all of which are challenging for linear models to describe accurately. Consequently, there has been a move toward more advanced approaches that offer increased flexibility and precision in modeling complex phenomena.

Artificial Neural Networks (ANNs) represent a substantial advancement in modeling complex physical and chemical systems. Emulating the neural structures of biological organisms, ANNs handle non-linear relationships and can adapt in real-time to new data. These networks excel at tasks like pattern recognition, classification, and prediction, making them invaluable for analyzing intricate system behaviors. ANNs effectively address issues such as non-linear interactions, high-dimensional data, and complex correlations, which are often difficult for traditional models to manage. Despite their strengths, ANNs have challenges related to model interpretability and generalization beyond their training data, yet these are minor compared to their significant power[67, 68].

The utilization of ANNs in physics and chemistry has become more popular due to their capability to model complex systems without the need for rigid assumptions on underlying processes. Networks such as the Multi-Layer Perceptron (MLP) have been successfully used in tasks including predicting molecular properties and simulating material behaviors. Comparative studies have demonstrated that ANNs can outperform traditional models in capturing sophisticated system dynamics and adapting to evolving data.

2.4.4 Structure of a Neural Network

Artificial Neural Networks (ANNs) are computational models inspired by the human brain's network of neurons. They consist of layers of interconnected units called neurons or perceptrons. The primary structure of an ANN includes an input layer, one or more hidden layers, and an output layer. Each neuron processes incoming signals using a weighted sum and an activation function to introduce non-linearity into the network. The building block of these networks is the perceptron, mathematically represented as[68]:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{2.38}$$

where x_i are the input features, w_i are the weights associated with each input, b is the bias term, and f is the activation function. The activation function, such as the sigmoid $f(x) = \frac{1}{1+e^{-x}}$, tanh $f(x) = \tanh(x)$, or ReLU $f(x) = \max(0, x)$, introduces non-linear properties to the network, enabling it to learn complex relationships in data.

A neural network becomes a "deep neural network" when it has multiple hidden layers between the input and output. The depth of the network allows it to capture high-level abstractions in data. For a network with L layers, the output of a neuron in the l-th layer can be represented as:

$$h_j^{(l)} = f\left(\sum_{i=1}^{n} w_{ij}^{(l)} h_i^{(l-1)} + b_j^{(l)}\right), \tag{2.39}$$

where $h_j^{(l)}$ is the output of the j-th neuron in the l-th layer, $w_{ij}^{(l)}$ are the weights connecting the neurons from layer $l-1$ to layer l, $h_i^{(l-1)}$ are the inputs from the previous layer, and $b_j^{(l)}$ is the bias term for the current layer.

In neural network training, the objective is to minimize the error between the predicted output and the actual output using a loss function. For example, the mean squared error (MSE) for regression tasks is given by:

$$E = \frac{1}{2}\sum_{k=1}^{m}(y_k - \hat{y}_k)^2, \tag{2.40}$$

where y_k is the true output, $\hat{y}_k$ is the predicted output, and m is the number of training examples.

Considering today's standards for training neural networks, we inevitably discuss backpropagation. In this algorithm, network weights are updated by propagating the error backwards, normally utilizing the gradient descent method to minimize the loss function but not limited to it. The weight update rule for each layer l is [67, 69]:

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \eta \frac{\partial E}{\partial w_{ij}^{(l)}}, \tag{2.41}$$

where η is the learning rate, and $\frac{\partial E}{\partial w_{ij}^{(l)}}$ is the partial derivative of the error with respect to the weight $w_{ij}^{(l)}$. This derivative, computed during backpropagation, indicates how much a change in each weight affects the overall error, guiding the optimization process.

Neural networks come in different architectures suited for various tasks. Feedforward Neural Networks (FNN) are the simplest, with information moving in one direction from input to output,

which makes training straightforward. Convolutional Neural Networks (CNN) are commonly used for image processing; they use filters to detect spatial features, followed by pooling layers to reduce data size, and end with fully connected layers for classification but can also be used in regression. Recurrent Neural Networks (RNN) are designed for sequential data, as they have loops that help them remember previous inputs, making them useful for time-series analysis and natural language processing. Lastly, Generative Adversarial Networks (GANs) involve two networks, a generator and a discriminator, that compete to improve their output, often used for image creation and data generation[70]. None of these architectures are exclusive to the uses mentioned here, but they represent the most common tasks for each type. It is possible to experiment with different architectures and parameters until a satisfactory result is achieved.

The optimisation of deep networks requires advanced techniques to overcome issues like vanishing gradients and overfitting. Methods such as dropout, batch normalisation, and adaptive learning rates (e.g., Adam) are employed to enhance learning and generalisation. The flexibility and adaptability of neural networks make them powerful tools in modelling complex physical and chemical systems, allowing for data-driven, non-linear approaches to solving problems that traditional models struggle with [68].

3 Methodology

It is essential that neural network-based force fields are trained on data that comprehensively covers the chemical configuration space of interest. The training dataset should display a wide range of chemical compositions, including diverse molecules and atomic connections, in various scenarios involving both bonded and non-bonded interactions, as well as different configurations. Since the primary goal of this thesis is to build a neural network model to supplant the classical force field in our system, obtaining top-quality data with these characteristics is vital. The overall method to achieve this is demonstrated in the subsequent image.

Note: An explanation of the coding and functioning of the software used is beyond the scope of this thesis. What is essential is that these tools employ the techniques, models, and approximations we have already reviewed in the background section. The code generated during this research process can be found in the following GitHub repository: https://github.com/MiguelChavez98/QM-final

3.0.1 Forcefield and Potential Model for Quin-PF6 System

Building on chemical intuition and past successes, modern force fields typically follow a functional form that extends the original AMBER force field model. This approach divides the total energy into bonded interactions (bond, angle, torsion) and non-bonded interactions (long-range electrostatics).[27]

For the molecular dynamics simulations of the Quin-PF6 system, this traditional approach to the forcefield was selected in combination with the Buckingham potential for the non-bonded interactions to better capture the nature of the system. Due to the ionic characteristics of Quin-PF6, which strongly influence both short- and long-range interactions, it was necessary to use this combination of potential energy terms to achieve accurate results. The Buckingham potential + electrostatics, given by the equation below, was chosen to replace the Lennard-Jones potential that is widely use in most systems for the treatment of interatomic interactions:

$$E_{jk} = B\exp\left(-Cr_{jk}\right) - Ar_{jk}^{-6} + \frac{q_j q_k}{r_{jk}} \tag{3.1}$$

This potential better accounts for the complex interplay between the repulsive and attractive forces within the ionic framework of Quin-PF6. The ionic interactions play a crucial role in the behavior of the system, driving both the short-range repulsive forces and the long-range electrostatic interactions. Taking these factors into account, the total potential energy of the system will be expressed as:

$$\begin{aligned} E = & \sum_{ij\in\text{Bonds}} k_r\left(r_{ij}-r_0\right)^2 + \sum_{ijk\in\text{Angles}} k_\theta\left(\theta_{ijk}-\theta_0\right)^2 \\ & + \sum_{ijkl\in\text{Torsions}} V_n[1+\cos(n\varphi_{ijkl}-\gamma)] \\ & + \sum_{jk\in\text{Buckingham + Elst}} \left(B\exp\left(-Cr_{jk}\right) - Ar_{jk}^{-6} + q_j q_k r_{jk}^{-1}\right) \end{aligned} \tag{3.2}$$

This combination allows for a more realistic representation of both the short-range repulsive and long-range electrostatic interactions typical of ionic systems, especially those with significant barocaloric effects like Quin-PF6. The intrinsic ionic structure and rotational freedom of the quinuclidinium cation enhance both configurational entropy and phase transitions, which are key to the barocaloric properties of this material.

3.1 Data Generation for forcefield Fitting

To accurately fit a forcefield for a specific system, it is essential to have high-quality data that reflect the physical behavior of the system. There are two primary approaches for generating such data. The first approach involves using experimental data from reliable sources, such as spectroscopic studies, calorimetric measurements, or X-ray diffraction. Experimental data are obtained using high-precision equipment and well-characterized materials, ensuring their reliability.

The second approach, which has gained popularity in recent years, involves avoiding real-world experiments, thereby saving on cost, energy, and time. These are the so-called ab initio methods, which do not rely on experimental data but instead use theoretical calculations grounded in quantum mechanics. At the heart of these methods is the Schrödinger equation, which theoretically provides exact results for a system. However, solving the Schrödinger equation exactly for most systems is computationally prohibitive. To address this, numerous optimisation techniques have been developed, each making compromises in one area to excel in another, as illustrated in the following figure.

The core of this project involves using some of these reliable computational methods to fit the forcefield described in the previous section. Once a system-specific forcefield has been fitted, it can be applied in molecular dynamics (MD) simulations to generate high-quality data on a wide range of configurations and their associated energies. This data is crucial for the subsequent neural network training phase.

Intermolecular forces determination

Intermolecular forces are described by the Buckingham potential and an electrostatic term. The former accounts for short-range interactions, while the latter represents long-range interaction energies. Although this formulation includes various intermolecular interactions such as electrostatic, exchange, induction, and dispersion, it is possible to isolate the electrostatic term, as it is the only component directly influenced by the system's charges. By separating the electrostatic contribution, the force field can be fitted more accurately to the remaining terms. If the electrostatic term is accurately modeled, this allows for more precise determination of the parameters A, B, and C, as well as the electrostatic values. This approach is particularly important given that the system is dominated by electrostatic interactions due to its ionic nature and the mobility of these ions [71].

$$E_{jk} = B\exp\left(-Cr_{jk}\right) - Ar_{jk}^{-6} + q_j q_k r_{jk}^{-1} \tag{3.3}$$

Considering this, a more efficient strategy for fitting the remaining terms of the potential (such as the A, B, and C parameters of the Buckingham potential) involves handling the electrostatic term separately. For the total energy calculations, the quantum chemistry software *Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry* [72] was employed, with the SAPT2 method chosen to compute the interatomic interaction energies. Although it is possible to try to fit all potential terms at once using the total energy and somehow assign charges to comply with the values in the optimisation, this method is inefficient and may fail to reach the minimum energy due to the high

dimensionality and complexity of the parameter space, or may assign highly unrealistic charges. As this is a key part of the model, I cannot allow this.

To address this, the software *CamCASP 7.2* was used for calculating the partial charges, leveraging its specialized capabilities for calculating interaction energies in molecular pairs [73]. This software is particularly well-suited for charge calculations due to its implementation of the ISA (Iterated Stockholder Atom) method [74], which provides high accuracy and computational efficiency by partitioning molecular electron densities into atoms with intuitively correct shapes and charges.

The workflow for this part is as follows: first, the total energy, incorporating all interatomic interactions, was calculated using Psi4 with the SAPT2 method (discussed in the background section). Simultaneously, CamCASP was employed to compute the charges with high accuracy. Using these precise charges, the electrostatic term was calculated and subtracted from the total energy. The remaining energy was then used to fit the A, B, and C parameters of the Buckingham potential, which enabled a more efficient and precise optimization of the potential terms and provided a better starting point for parameter fitting. This approach reduces computational time and increases accuracy, particularly in the electrostatic interactions.

All the programs require a description of the system, with specific formats described in their respective sections. However, as a general rule, input files or scripts for calculations typically use XYZ coordinates and atom types to generate results. The coordinates for the Quin-PF6 system were extracted from a crystallography file (.cif) and subsequently converted to XYZ format using the molecular visualisation software Mercury (version 4.0)[75].

Electrostatic interactions

There are various methods to calculate electrostatic interactions and charges, including Mulliken Population Analysis, which assigns charges based on the overlap of atomic orbitals but can lead to inconsistent results [76]. Natural Population Analysis (NPA) provides a more consistent charge assignment using natural orbitals[77]. The Iterated Stockholder Atom (ISA) method offers highly accurate atomic charges by partitioning molecular electron densities into atoms with intuitive shapes and charges. This method iteratively refines these charges to minimize discrepancies between calculated and target densities, making it a robust option for systems dominated by electrostatic interactions. The effectiveness of the ISA method is highlighted in the work by Misquitta and Stone (2019) [73].

The *Iterated Stockholder Atoms (ISA)* method is implemented in **CamCASP 7.2**. To perform these calculations, **CamCASP** utilizes the **Psi4** quantum chemistry software to carry out **Density Functional Theory (DFT)** calculations, generate basis sets, and extract relevant molecular information. CamCASP also requires input files to define the specific parameters and configurations for the calculations.

A critical step in this process is obtaining the **Highest Occupied Molecular Orbital (HOMO)** eigenvalue and the **Ionization Potential (IP)**, or alternatively, the **AC-shift** value. To determine these properties, I employed **Psi4**, using the **PBE0** exchange-correlation functional and the **aug-cc-pVTZ** basis set, alongside a **hybrid CHF+ALDA** response kernel. The built-in functions within Psi4 facilitated the extraction of the HOMO eigenvalue and the Ionization Potential. These calculations were executed via a Python script interfacing with the Psi4 API, and run on a high-performance cluster equipped with **504 GB of RAM and 64 cores**.

Once the HOMO eigenvalue and the Ionization Potential were computed, I generated the input file required for running the ISA module in CamCASP. This input file was executed on the same cluster. An example of the input file code is provided in the appendices. The resulting output file contains

the atomic charges calculated via the ISA method.

The ISA method utilizes an **auxiliary basis set**. While CamCASP can perform a **multipole expansion** to enhance the description of electrostatic interactions and retain the **Williams potential**, I opted for a simplified model using **normal point charges** on each atom to streamline the calculations.

Total Intermolecular Energy Calculation Using SAPT2

Symmetry Adapted Perturbation Theory - 2 (SAPT2) was employed to calculate intermolecular interaction energies. As mentioned in the background section the designation "2" refers to the second-order corrections applied within this theoretical framework to enhance precision. SAPT2 decomposes the total interaction energy into physically meaningful contributions, including electrostatics, exchange, induction, and dispersion. This decomposition allows a detailed analysis of the different forces contributing to the total energy, making SAPT2 particularly valuable for systems where a thorough understanding of these interactions is needed.

The method was implemented using Psi4 and run on the same computational cluster mentioned earlier. For our calculations, we selected a diverse set of molecular configurations in both random and ordered sequences, ensuring a comprehensive representation of the system's interaction landscape. The high precision of SAPT2 ensures that the resulting energies provide a reliable reference set for forcefield fitting.

The fitting process followed a structured approach: in order to determine the A, B and C constants of the Williams potential for two interacting molecules, each type of atom in the molecule must have its own set of values for these constants. Additionally, combination rules were applied when different types of atoms interact, which were essential for implementing the forcefield in code and in the MD simulations. The combination rules are given by:

$$A_{jk} = (A_{jj}A_{kk})^{1/2} \quad B_{jk} = (B_{jj}B_{kk})^{1/2} \quad C_{jk} = 0.5\,(C_{jj} + C_{kk}) \tag{3.4}$$

Initially, since PF6 does not have pre-existing values for A, B and C the parameters from Quin atoms were fixed, and the parameters for PF6 were optimized to match the intermolecular energy between Quin and PF6, subtracting the electrostatic energy using pre-calculated charge distributions. This provided an initial set of values for PF6.

Next, a second optimization step was conducted, allowing the constants for both Quin and PF6 to vary, leading to an improved fit. The final step involved a refined optimization with quadratic penalties applied when parameter deviations exceeded 15%, ensuring the results remained physically meaningful. The penalty factor was set to 2000 based on empirical tests, and the quadratic deviation was added as a penalty to prevent overfitting.

To generate the molecular configurations for SAPT2 calculations, two distinct sets were used. For the disordered configurations, a custom script was developed to randomly place points on the surface of a sphere with random rotations, at distances of 5, 6, 7, and 9 Å. Each sphere contained 20 points, and the distances were measured from the geometric center of the molecule. These distances were chosen based on preliminary test runs, which indicated they provided a good range of energies for the fitting process.

For the ordered configurations, rotations were applied every 10 degrees over a 180-degree sweep along the Y-axis, at distances of 6, 7, and 8 Å. In these configurations, PF6 was kept fixed in the same orientation as in the original ".cif" file, capturing structured interactions with Quin.

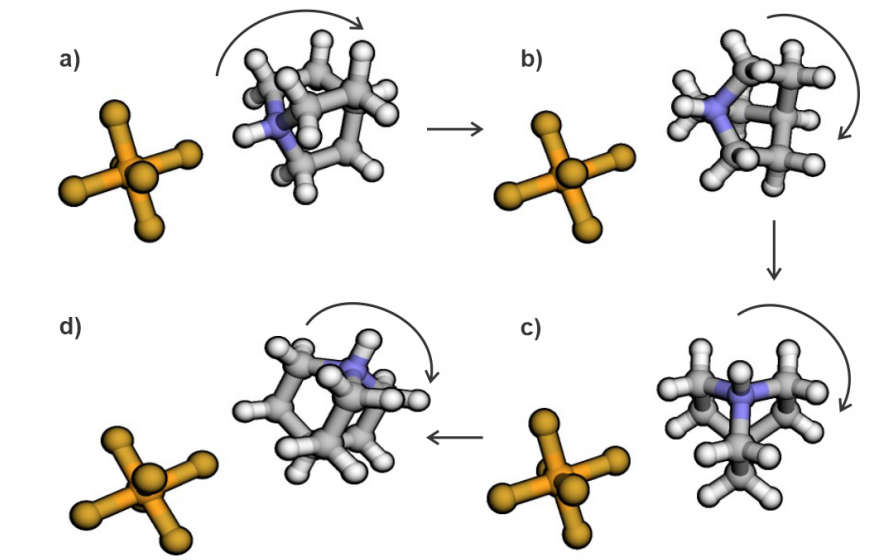


Figure 3.1: Rotations of Quin-PF6 to calculate new energies in various configurations

In total, this resulted in 137 configurations (80 disordered and 57 ordered), providing a robust dataset for SAPT2 energy calculations. These configurations allowed for a detailed exploration of the interaction landscape, which was critical for forcefield fitting.

Once the configurations were established, SAPT2 calculations were performed on the computational cluster, utilizing 16 cores,with a parallelized setup optimized for long computations. SAPT2 was chosen due to its ability to provide accurate decompositions of intermolecular interactions into electrostatics, exchange, induction, and dispersion contributions, which are essential for understanding the physical nature of the system. The hybrid functional PBE0 was used for exchange and correlation, paired with the aug-cc-pVTZ basis set. This functional provides a balanced approach between Hartree-Fock exchange and density functional theory correlation, which is essential for precisely capturing the interaction energies in not so common systems such as Quin-PF6. The aug-cc-pVTZ basis set provides an adequate handling of electron correlation while keeping the computational expense within manageable limits[48].

3.1.1 Forcefield Optimization

The optimization of the forcefield was carried out using the Nelder-Mead algorithm, a simplex-based optimization method that does not require the calculation of derivatives. This method is particularly useful in scenarios where functions are nonlinear and do not lend themselves to derivative-based approaches, making it suitable for complex optimization problems[78].

During the optimization process, attempts were also made using gradient descent methods, specifically employing the Adam optimizer, which is known for its efficiency in handling large datasets and sparse gradients. The Adam optimizer updates parameters based on the first and second moments of the gradients, allowing for adaptive learning rates [79].

Additionally, optimization using genetic algorithms was explored as a potential alternative. Genetic algorithms are inspired by the process of natural selection and employ operations such as selection, crossover, and mutation to evolve solutions toward optimality. However, after careful evaluation, I decided to proceed with the Nelder-Mead algorithm due to its effectiveness and simplicity for the problem at hand [80]

The implementation of Nelder-Mead from the SciPy library was used for the calculations. The default parameters of this algorithm, which are effective for most test functions, were employed to

ensure a robust and reliable optimization approach. SciPy is widely recognized for its efficiency and accuracy in scientific computations and is a commonly used tool within the research community[81].

In the implementation of the Nelder-Mead algorithm using the SciPy library, several default parameters were utilized to ensure effective optimization. The tolerances for convergence were set to $1e-4$ for both the solution ($xatol$) and the objective function value ($fatol$). The maximum number of iterations was specified as 2000, allowing sufficient computational steps for convergence. Additionally, the option to display output messages during the optimization process was disabled by default, ensuring a cleaner execution without unnecessary console output. These parameters were chosen to balance the need for accuracy and computational efficiency during the forcefield optimization process.

The initial data for the optimization was derived from the previous commented method SAPT2, where interaction energies between the molecules were calculated. Using these values as a reference, the parameters of the Williams potential—A, B, and C—were optimized to minimize the discrepancy between the calculated energies and the experimentally obtained ones. This optimization process was crucial for obtaining a fitted forcefield parameters mentioned in the past section that was utilized in subsequent molecular dynamics simulations.

3.1.2 Application of the Optimized Forcefield in Molecular Dynamics

In the initial stage of the molecular simulation study, the structural data of the Quin-PF6 unit cell were prepared and manipulated. Originally configured as a 1x1 cell with a single Quin-PF6 molecule, the data were converted into PDB (Protein Data Bank) format, which is widely used for storing three-dimensional structures of biological molecules and molecular assemblies. Using the molecular visualization software Mercury and a custom Python script, the cell was expanded to a 2x2 supercell configuration, ensuring precise placement of molecules for further simulations.

OpenMM 7.7 [42] was chosen for its high performance and customizable environment, which was critical for integrating the Buckingham potential. The software's ability to define custom intermolecular energy expressions enabled accurate modeling of Quin-PF6 interactions, essential due to the ionic interactions between the PF6 anion and Quin cation.

The simulation environment was set up by loading the 2x2 formatted structural data into OpenMM. A system object was initialized, and particles were added according to types and masses from a predefined dictionary. Harmonic bond and angle forces were applied by iterating over the molecular topology, using parameters from dictionaries tailored for each bond and angle.

A custom nonbonded force was implemented to account for specific intermolecular forces, especially the Buckingham potential. This included parameters for exponential repulsion and attractive dispersion, aligned with the unique properties of Quin-PF6. This was vital for modeling interactions effectively over various distances.

In the technical configuration of the simulations, multiple simulation runs were employed with a friction range of 1 to 4 picoseconds^{-1}, allowing variation in system damping to examine different kinetic relaxation behaviors. Temperature ranges from 200 to 500 Kelvin were explored to ensure that the critical phase transition, documented between 300 and 400 Kelvin for barocaloric materials, was within the simulated conditions. For system stabilization, an NVT (Canonical ensemble) was utilized, ideal for controlling temperature while keeping the number of particles and volume constant. An NPT (Isothermal-Isobaric ensemble) followed for adjusting volume and pressure along with temperature, mimicking real-world thermodynamic experiment conditions. The simulations were configured to run 25,000 steps, with the center of mass motion disabled to maintain system stability. Total energies,

potential energies, and kinetic energies, along with trajectories, were recorded and stored in DCD files, with snapshots taken every 20 steps to provide detailed temporal evolution of the system.

3.1.3 Software for data analysis and visualisation

To visualize and analyze the molecular dynamics (MD) trajectories, several tools and libraries were employed. The MD data were saved in DCD but converte to PDB, a standard format used to store atomic coordinates over time. This format facilitated both the visualization and subsequent analysis of the trajectories.

For visualization, the `py3Dmol` library was utilized to render molecular structures interactively. The PDB files generated from the simulations were read, and the trajectory was visualized frame by frame. Each frame was loaded as an individual model, allowing for a detailed inspection of the atomic movements over the course of the simulation. The unit cell was applied to the visualization to ensure that periodic boundary conditions were maintained, and the dynamic behavior of the molecules was animated. Specific rendering styles, such as sticks and spheres, were used to enhance the clarity of the molecular structure during the animation.

Additionally, the `MDAnalysis` library [82] was employed to process the molecular trajectories and extract specific atomic positions for further analysis. The orientations of selected molecular bonds were tracked across the trajectory. For each frame, bond vectors between atoms of interest were computed, and their polar and azimuthal angles were determined. The polar angle θ represents the angle between the bond vector and the z-axis, while the azimuthal angle ϕ corresponds to the bond orientation in the x-y plane. These angular values were recorded at each timestep to facilitate subsequent analysis.

3.1.4 Conversion to Descriptors and Neural Network Application

For the analysis of molecular dynamics simulations, trajectories reflecting the expected behavior of the system across various temperatures were selected. Each simulation, stored in DCD format, contains 12,500 frames, capturing the dynamics under different entropic conditions. These frames exhibit distinct molecular movements: outward-pointing bonds in low-entropy phases and more dispersed movements in high-entropy phases, making them ideal for detailed analysis.[83]

Descriptor Calculation

The descriptors were calculated using the Smooth Overlap of Atomic Positions (SOAP), which is particularly suitable for periodic or large systems. The SOAP parameters were set to include atomic species such as Carbon (C), Hydrogen (H), Nitrogen (N), Phosphorus (P), and Fluorine (F), with a cutoff radius r-cut of 5.0, maximum degree of radial basis functions n-max of 8, and maximum angular momentum quantum number l-max of 8. This configuration was chosen for its effectiveness in capturing the local chemical environment, utilizing an 'inner' average to aggregate the descriptors over the trajectory, thus providing a representative description of the system's state over time. [84]

Neural Network Architecture and Training

Given the complexity of the data, a straightforward neural network architecture was employed, focusing on regression to predict numerical outcomes related to the system's properties. The networks will consist in an input layer, and a changing number of hidden layers and neurons, and an output layer designed for regression. Each frame of the simulation was represented by approximately 3,700 descriptor values, which served as inputs to the network.

The dataset comprised a total of 12510 frames across 10 simulations. Each frame was independently fed into the network, which was trained to map the high-dimensional descriptor data to a scalar output reflecting a specific property of interest, such as the phase state or entropic level of the system.

Data Normalization and Network Training

Normalization of the data was deemed unnecessary due to the uniform scale of the descriptor values, which were within a similar range but varied enough to be distinguishable by the neural network. The network was trained using a simple split of the data, allocating 80% for training and 20% for validation. The training process focused on minimizing the prediction error, with adjustments made based on the validation set performance.

For performance metrics, R-squared, root mean squared error (RMSE) and mean absolute error(MAE) were used to evaluate the network's accuracy and predictive power. These metrics are standard for regression tasks, providing insight into the average error magnitude and the proportion of variance explained by the model.

4 Results and discussions

The results of this research can be divided into three main sections: the fitting of the forcefield and its ability to accurately reproduce the evaluated energy configurations, the correct simulation of the system to identify phase transitions and barocaloric behavior consistent with real-world observations, and finally, the fitting of the neural network to the forcefield and molecular dynamics data, assessing its capacity to reproduce the energies of the forcefield and its potential as a tool for generating and evaluate large quantities of data. Each section represents a significant milestone in the project, with experiments meticulously reproduced and carefully checked to ensure reliable outcomes in subsequent phases.

4.1 Force field Fitting and Energy Reproduction

The forcefield equation used for the Quin-PF6 system is:

$$\begin{aligned} E = & \sum_{ij \in \text{bonds}} k_r \left(r_{ij} - r_0\right)^2 + \sum_{ijk \in \text{angles}} k_\theta \left(\theta_{ijk} - \theta_0\right)^2 \\ & + \sum_{ijkl \in \text{torsions}} V_n[1 + \cos(n\varphi_{ijkl} - \gamma)] \\ & + \sum_{jk \in \text{non-bonded}} \left(B \exp\left(-C r_{jk}\right) - A r_{jk}^{-6}\right) \\ & + \sum_{jk \in \text{electrostatics}} \frac{q_j q_k}{r_{jk}} \end{aligned}$$

For the bonds, angles, and dihedrals, parameters from the OPLS-AA forcefield were used [30]. Therefore, the focus will now shift to the fitting of the non-bonded interactions, which are expressed by a combination of the Buckingham potential and electrostatic interactions as:

$$E_{jk} = B \exp\left(-C r_{jk}\right) - A r_{jk}^{-6} + \frac{q_j q_k}{r_{jk}}$$

After using the dataset consisting of 137 configurations and their associated energies, obtained through SAPT2 calculations as detailed in the methodology, the fitting results were obtained in the following order:

4.1.1 PF6-PF6 Optimization

The initial optimization was performed using 12 interaction data points for the PF6-PF6 system at varying distances, starting with an initial guess for the parameters of the Buckingham potential $\left(\text{kJ/mol}, \text{Å}\right)$. However, both the initial trials and subsequent optimizations using genetic algorithms did not yield satisfactory results, as the predicted interaction energies significantly deviated from expected values.

The parameters values used in the initial energy calculation are shown below:

Species	A	B	C
Fluorine (F)	11316.36	0.00	4.93
Phosphorus (P)	12195.91	903840.98	1.85

Table 4.1: Initial Buckingham parameters for PF6-PF6 interactions, highlighting suboptimal performance.

Generating the following results:

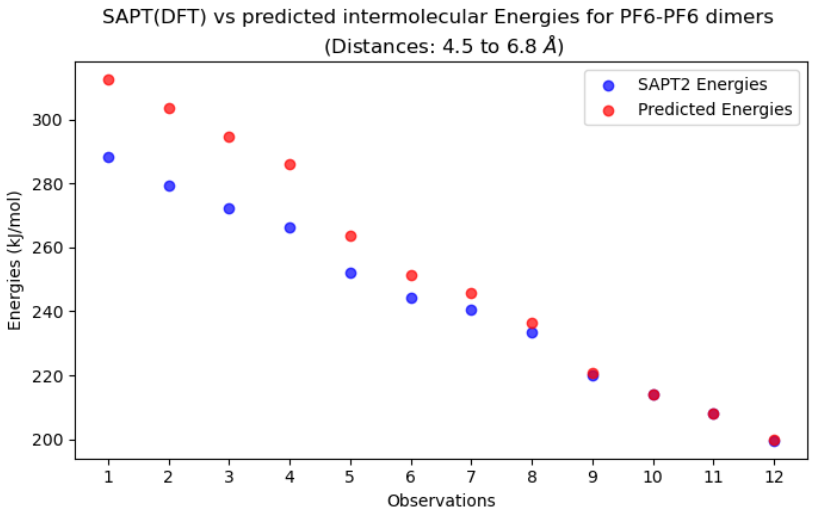


Figure 4.1: Comparison of intermolecular energies from SAPT(DFT) with the forcefield energies (Buckingham potential plus electrostatics) for PF6-PF6 interactions, in kJ/mol.

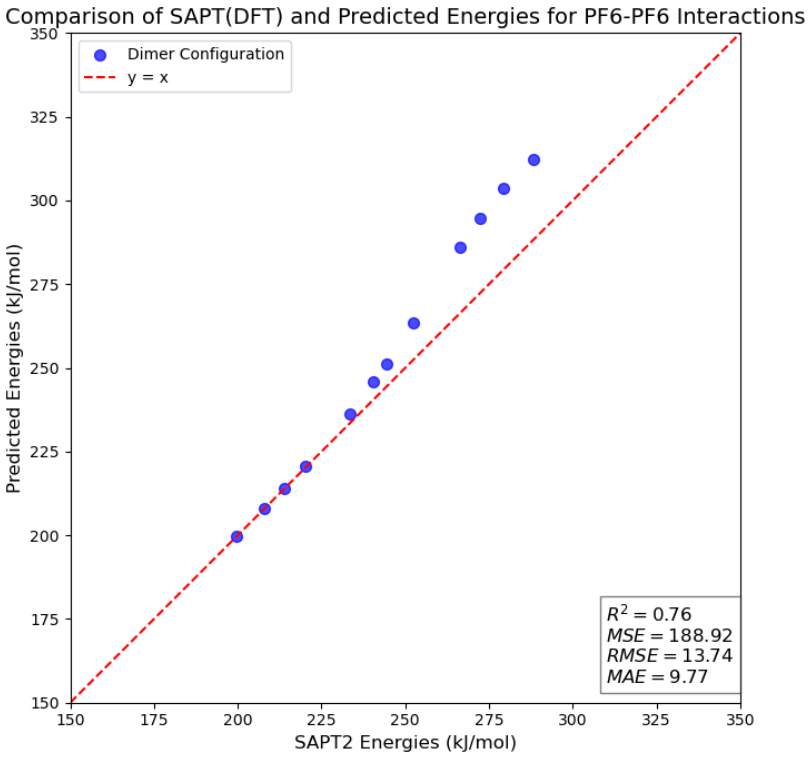


Figure 4.2: Parity plot comparing SAPT(DFT) and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.7597$, MSE = 188.92, RMSE = 13.74, MAE = 9.77, and $n = 12$.

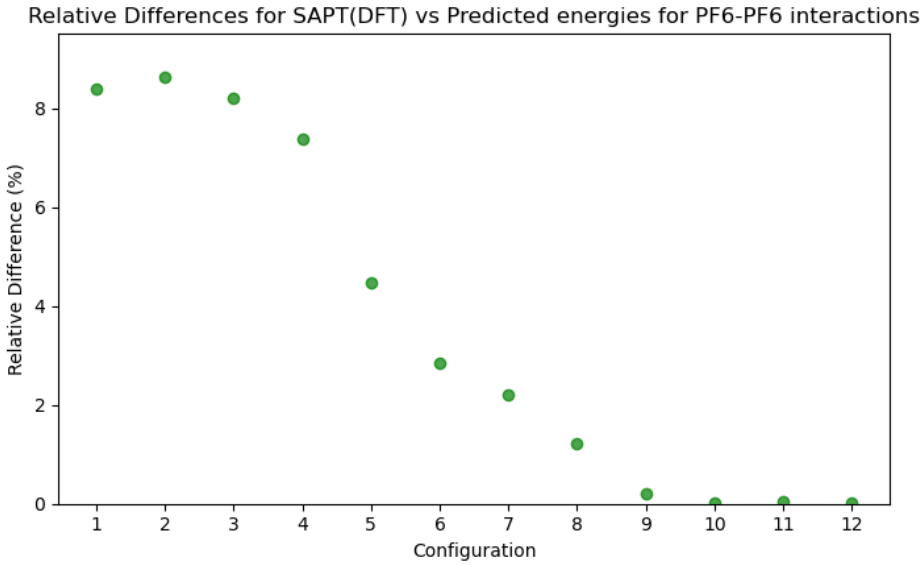


Figure 4.3: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT(DFT) energies for PF6-PF6 interactions.

The initial results did not meet the expectations for a well-performing forcefield, as evidenced by the scatter and parity plots. The correlation R^2 value was 0.7597, with errors exceeding 8% in some cases, leading to the exploration of alternative optimisation methods. These efforts resulted in a satisfactory fit of the interaction parameters using the Nelder-Mead algorithm. The final optimized values for parameters A, B, and C are presented below:

Species	A	B	C
Fluorine (F)	6289.67	218011.84	3.51
Phosphorus (P)	628.97	262227.46	2.88

Table 4.2: Optimized Buckingham parameters for PF6-PF6 interactions.

Following this optimization, the fitting results can be visualized in the following graphs:

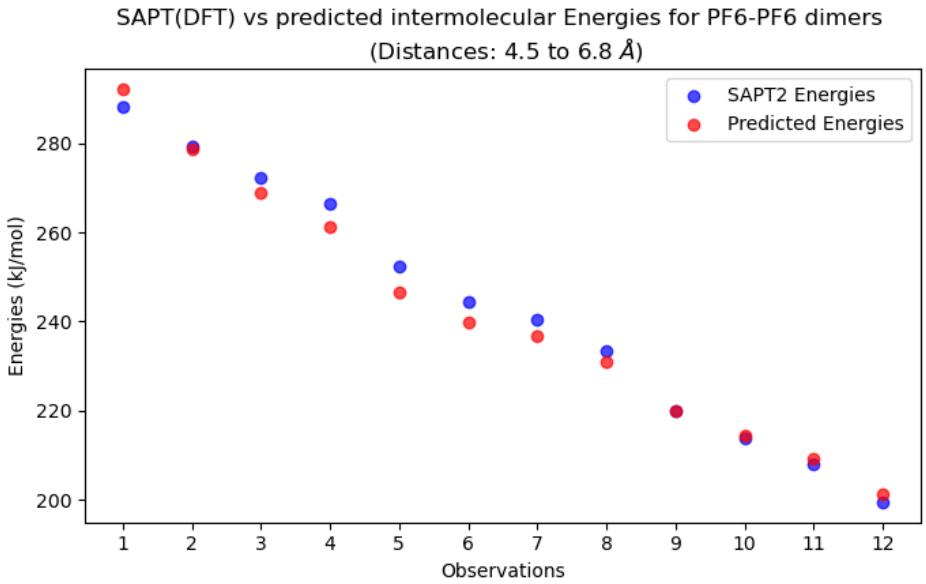


Figure 4.4: Comparison of intermolecular energies from SAPT(DFT) with the forcefield energies (Buckingham potential plus electrostatics) for PF6-PF6 interactions, in kJ/mol.

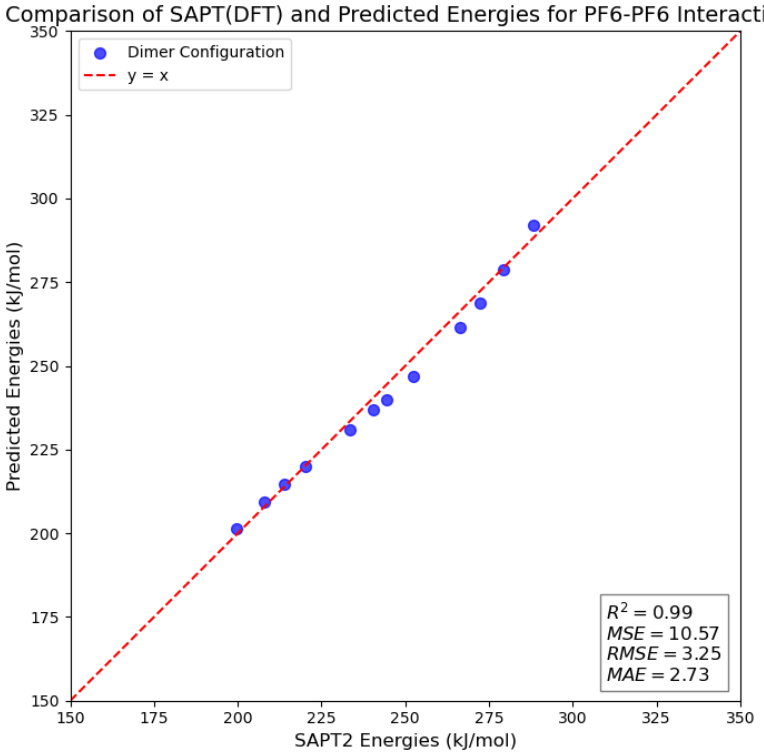


Figure 4.5: Parity plot comparing SAPT(DFT) and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.9865$, MSE = 10.57, RMSE = 3.25, MAE = 2.73, and $n = 12$.

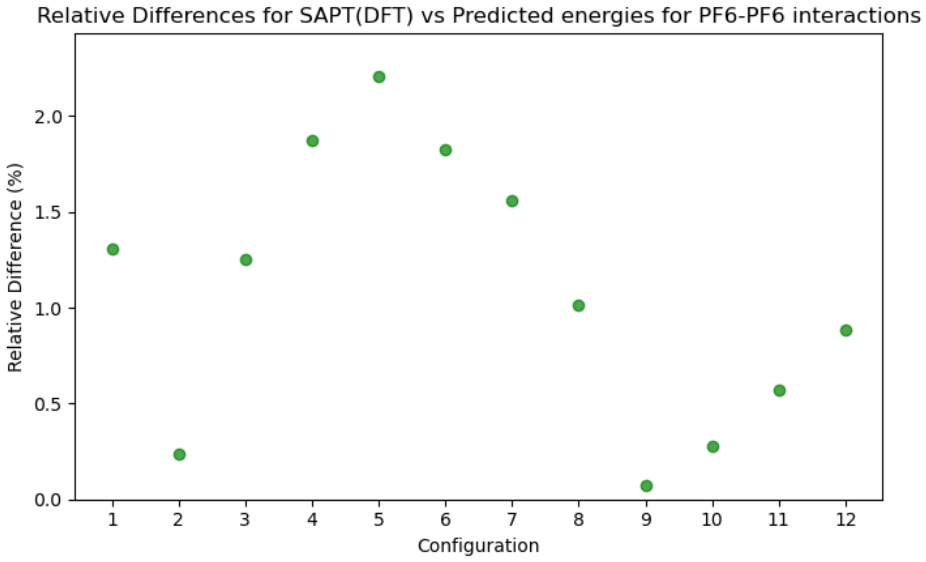


Figure 4.6: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT(DFT) energies for PF6-PF6 interactions.

The comparison between SAPT(DFT) and the intermolecular energies predictions shows good agreement, as illustrated by the scatter plot and parity plot. The R^2 value for the correlation is 0.9865, confirming the accuracy of the fitting, with all relative errors below 2.5%.

4.1.2 QUIN-QUIN Optimization

The second optimization was performed for the QUIN-QUIN dimer system using 64 configurations and energies from SAPT2, with initial parameters based on the values provided in Williams' paper [85], which describes the Buckingham potential parameters $(\mathrm{kJ/mol}, \mathrm{\AA})$ for various atom types.

The initial parameters for the Buckingham potential are:

Species	Parameter A	Parameter B	Parameter C
Carbon C(4)	978.36	131571.00	3.60
Hydrogen H(4)	0.00	764.90	3.56
Hydrogen H(1)	278.37	12680.00	3.56
Nitrogen N(3)	2376.55	191935.00	3.48

Table 4.3: Initial Buckingham parameters for QUIN-QUIN interactions based on Williams (2001).

The set of values derived from the Williams potential showed the following results, revealing inconsistencies in energy predictions and unexpected deviations, which suggest limitations in the potential's applicability to this system.

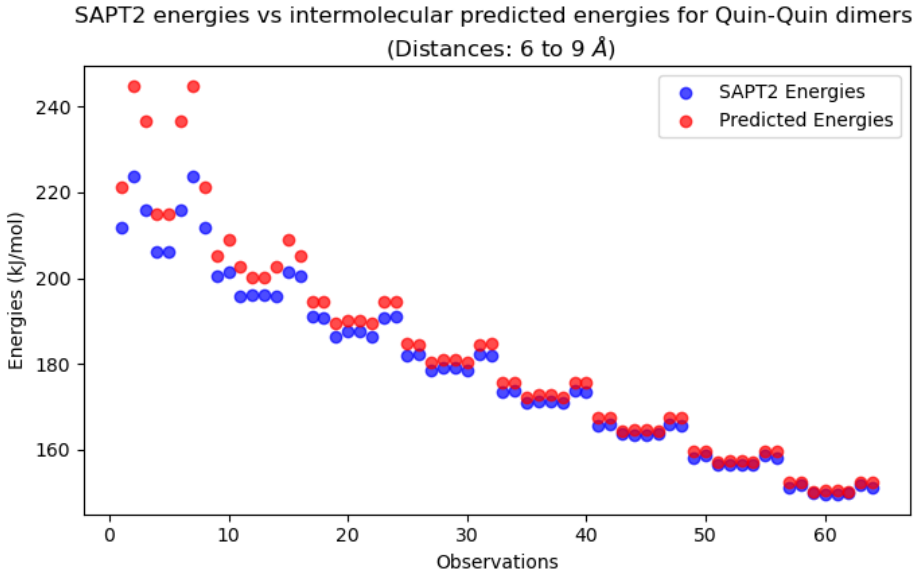


Figure 4.7: Comparison of intermolecular energies from SAPT2 with the forcefield energies (Buckingham potential plus electrostatics) for Quin-Quin interactions, in kJ/mol.

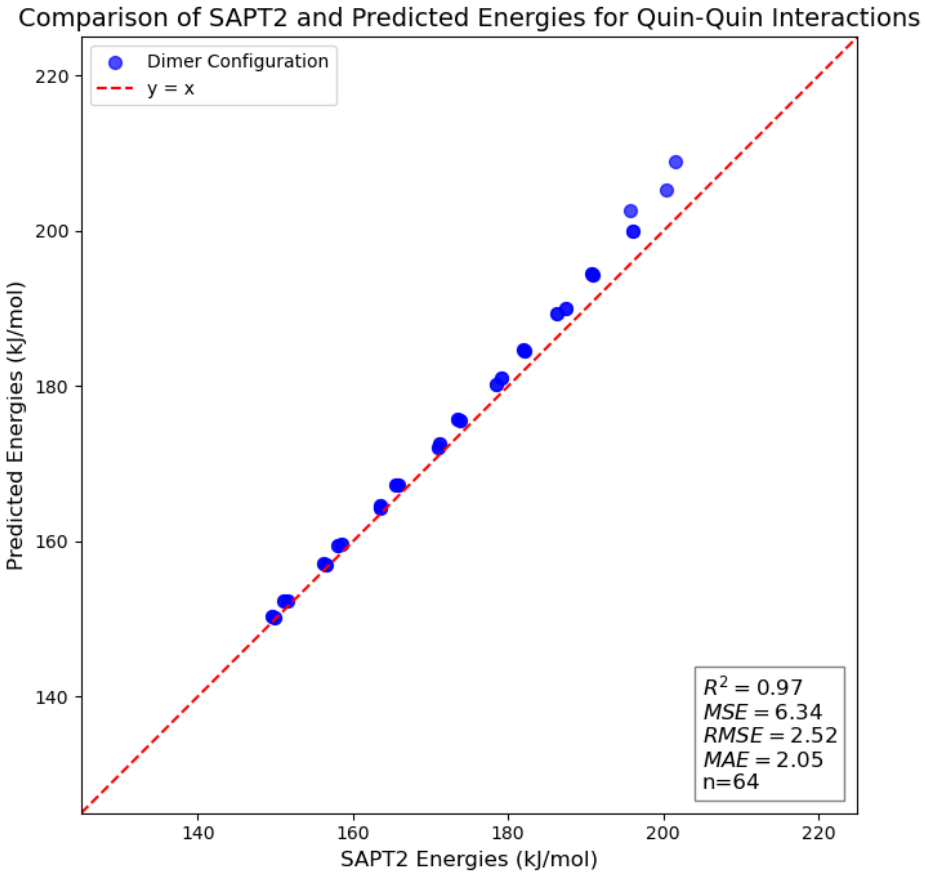


Figure 4.8: Parity plot comparing SAPT2 and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.9724$, MSE = 6.34, RMSE = 2.52, MAE = 2.05, and $n = 64$.

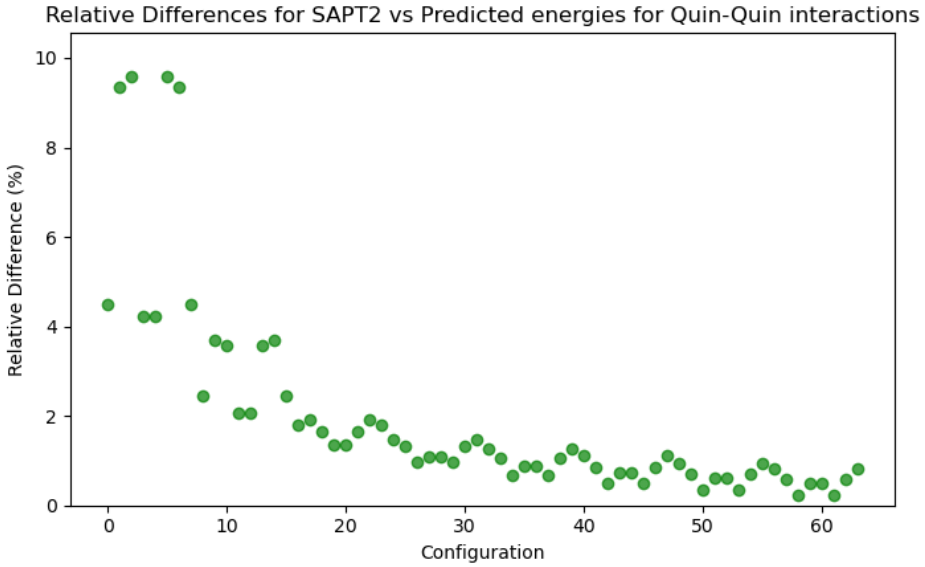


Figure 4.9: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT2 energies for Quin-Quin interactions.

The initial results did not meet expectations for a well-performing forcefield, as evidenced by the relative errors. Although the correlation R^2 value was 0.9724, which is a strong correlation coefficient,

some values exhibited errors exceeding 9.5%, leading to the exploration of alternative optimization methods. These efforts resulted in a satisfactory fit of the interaction parameters using the Nelder-Mead algorithm. The final optimized values for parameters A, B, and C are presented below:

Species	Parameter A	Parameter B	Parameter C
Carbon C(4)	1075.314	125229.826	3.889
Hydrogen H(4)	0.001	757.354	3.738
Hydrogen H(1)	250.547	11740.279	3.916
Nitrogen N(3)	2218.600	182190.172	3.183

Table 4.4: Optimized Buckingham parameters for QUIN-QUIN interactions.

The successful fitting results for QUIN-QUIN interactions can be better visualized in the following graphs:

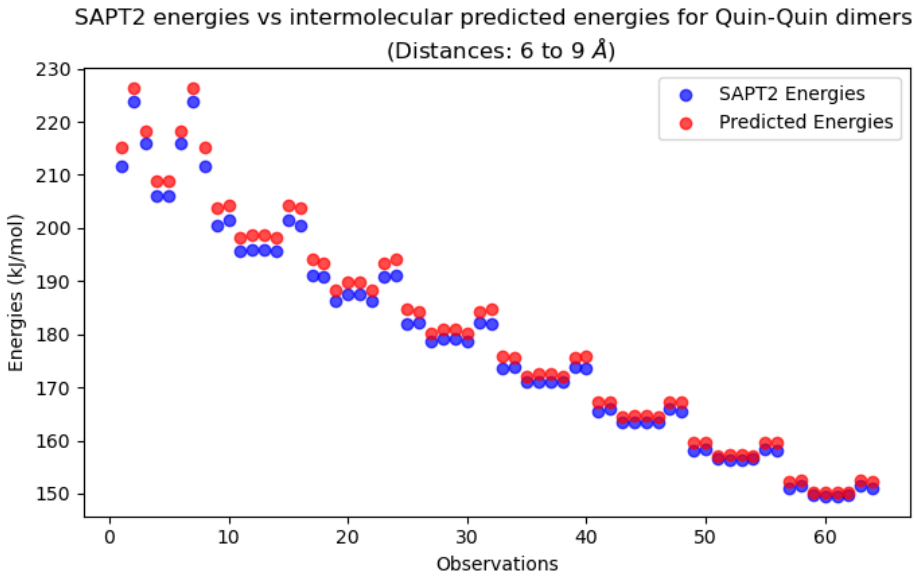


Figure 4.10: Scatter plot comparing SAPT2 energies with intermolecular predicted energies (Buckingham potential plus electrostatics) for QUIN-QUIN interactions, in kJ/mol.

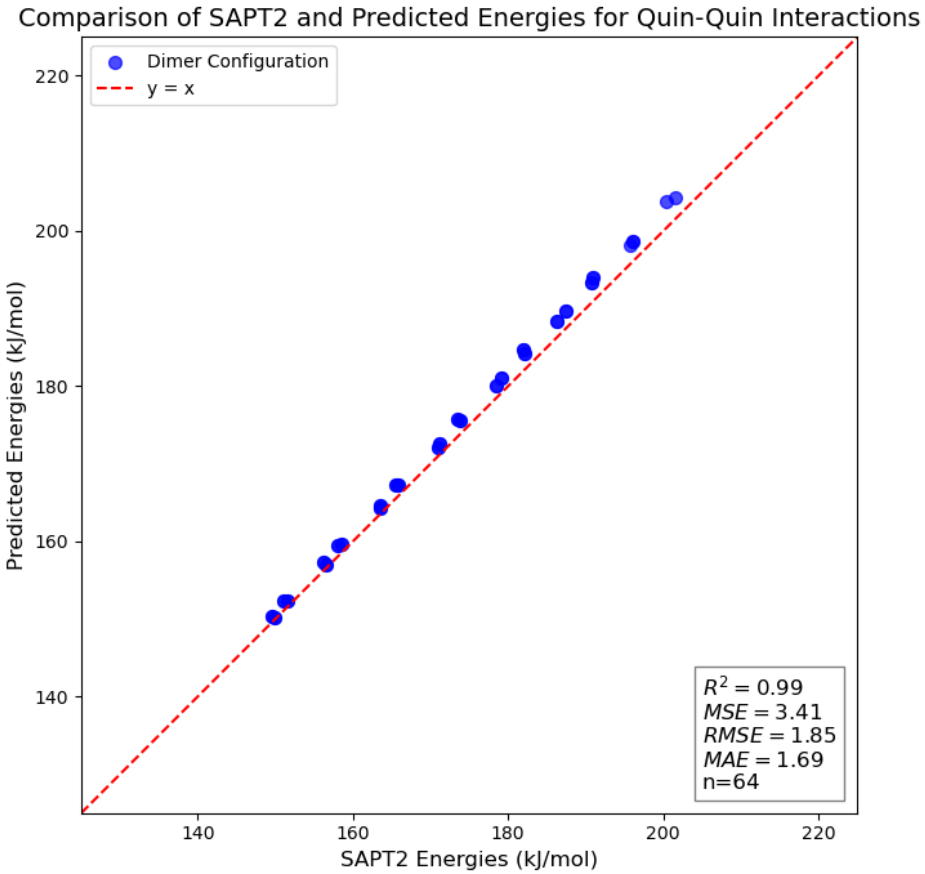


Figure 4.11: Parity plot comparing SAPT2 and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.9852$, MSE = 3.41, RMSE = 1.85, MAE = 1.69, and $n = 64$.

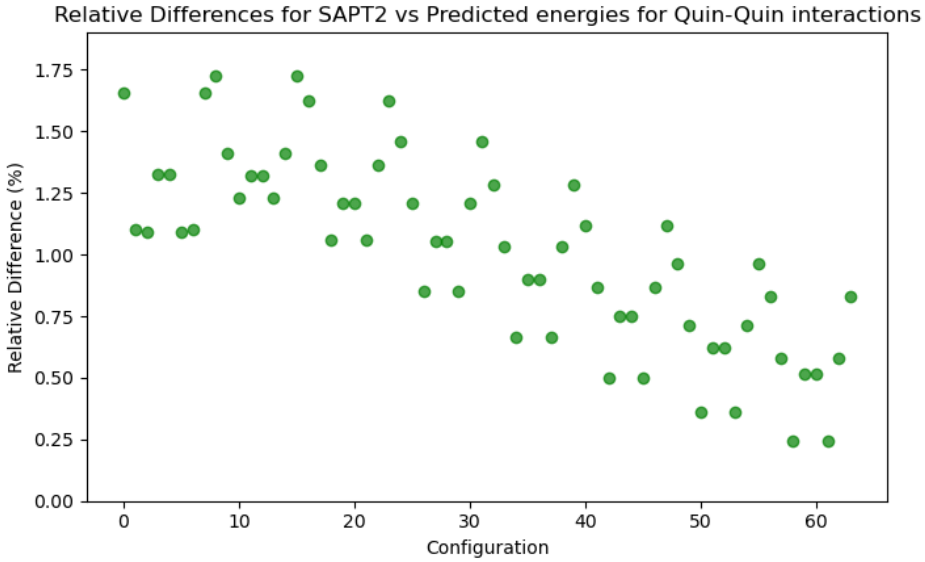


Figure 4.12: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT2 energies for Quin-Quin interactions.

The comparison between SAPT2 and the intermolecular energies predictions shows good agreement, as illustrated by the scatter plot and parity plot. The R^2 value for the correlation is 0.9986, confirming

the accuracy of the fitting, with all relative errors below 1.80% over all the different configurations.

4.1.3 QUIN-PF6 Optimization

Using the dataset of 137 configurations and their corresponding energies from SAPT2 calculations, the final optimization was focused on the interaction between QUIN and PF6. Initial parameters were taken from the previously optimized values for PF6-PF6 and QUIN-QUIN interactions. Although these parameters were fitted for the individual molecules, they showed a poor representation of intermolecular energies when using the combined parameters, as shown below:

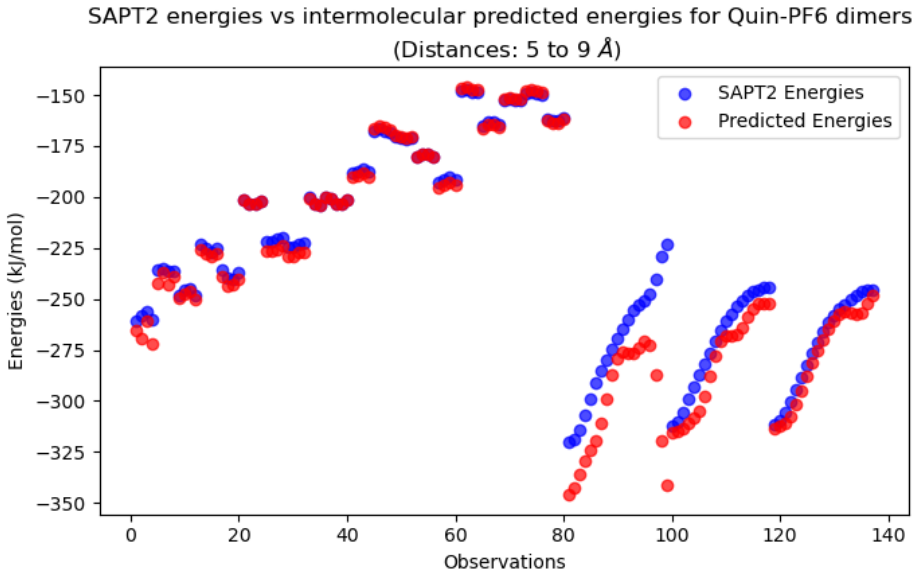


Figure 4.13: Scatter plot comparing SAPT2 energies with predicted intermolecular energies (Buckingham potential plus electrostatics) for QUIN-PF6 interactions, in kJ/mol. This set of configurations includes both randomly generated and systematically organized configurations.

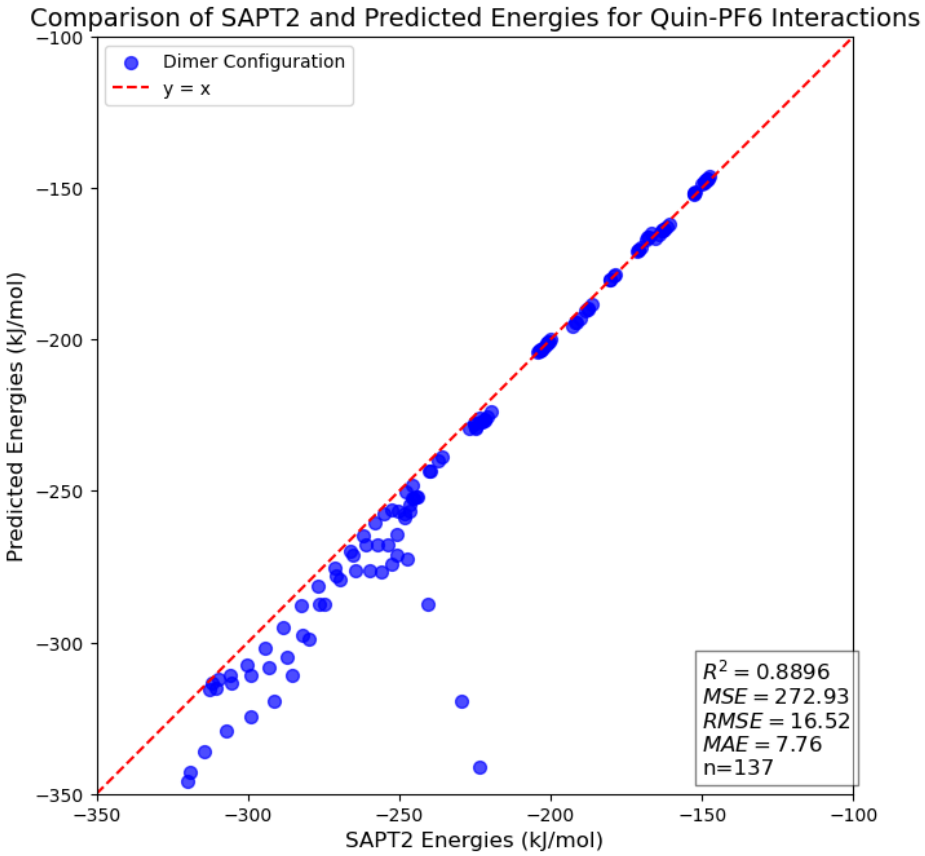


Figure 4.14: Parity plot comparing SAPT2 and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.8896$, MSE = 272.93, RMSE = 16.52, MAE = 7.76, and $n = 137$.

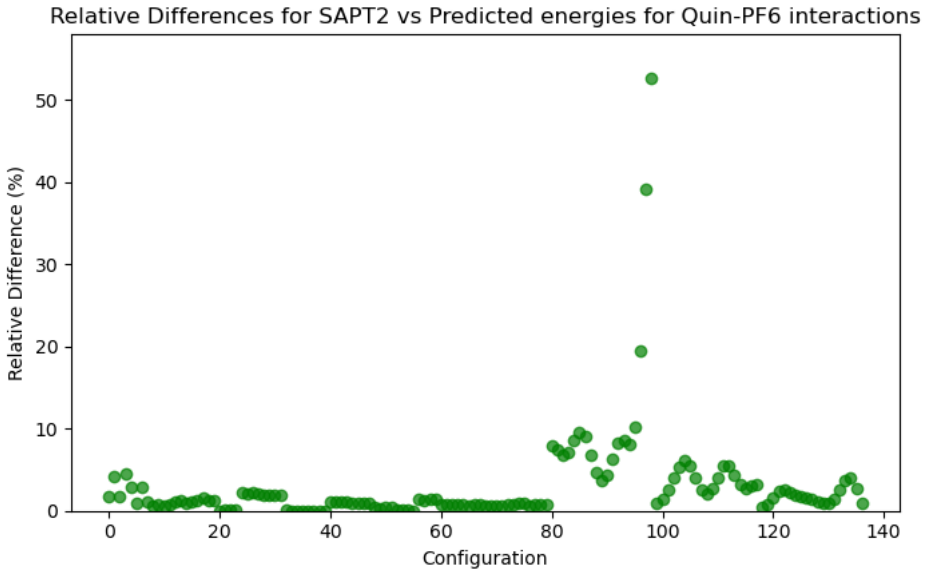


Figure 4.15: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT2 energies for Quin-PF6 interactions.

These results show very low performance when using the combined set of parameters through the combination rules, demonstrating the need for further optimisation by fitting all the parameters

simultaneously to better capture the real relationship between variables. To maintain physically realistic results, a penalty function was applied to prevent the parameters from deviating excessively during the fitting process to remain consistent with the previously established interactions. The penalty function allows for slight adjustments in the parameters to achieve an accurate fit while preserving the physical integrity of both the same-molecule interactions and the ion pair interactions (QUIN-PF6).

The final optimized parameters for QUIN-PF6 interactions are:

Species	A	B	C	Maximum Variation (%)
Fluorine (F)	6189.224	196313.208	3.798	11 %
Phosphorus (P)	572.188	273974.636	2.869	10 %
Carbon C(4)	967.866	129782.548	3.500	11.11 %
Hydrogen H(4)	0.000	732.504	4.112	10 %
Hydrogen H(1)	274.610	11125.934	3.815	9.6 %
Nitrogen N(3)	2173.827	200409.189	3.375	10 %

Table 4.5: Optimized Buckingham parameters for all species in QUIN-PF6 interactions, including the maximum variation observed in any parameter (A, B, or C).

The fitting results for QUIN-PF6 interactions can be better visualized in the following graphs:

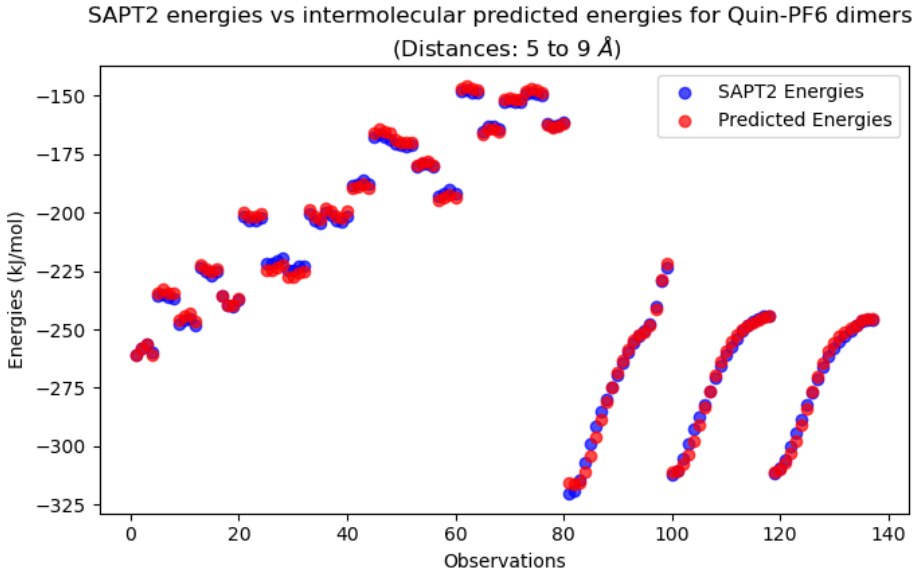


Figure 4.16: Scatter plot comparing SAPT2 energies with predicted intermolecular energies (Buckingham potential plus electrostatics) for QUIN-PF6 interactions, in kJ/mol. This set of configurations includes both randomly generated and systematically organized configurations.

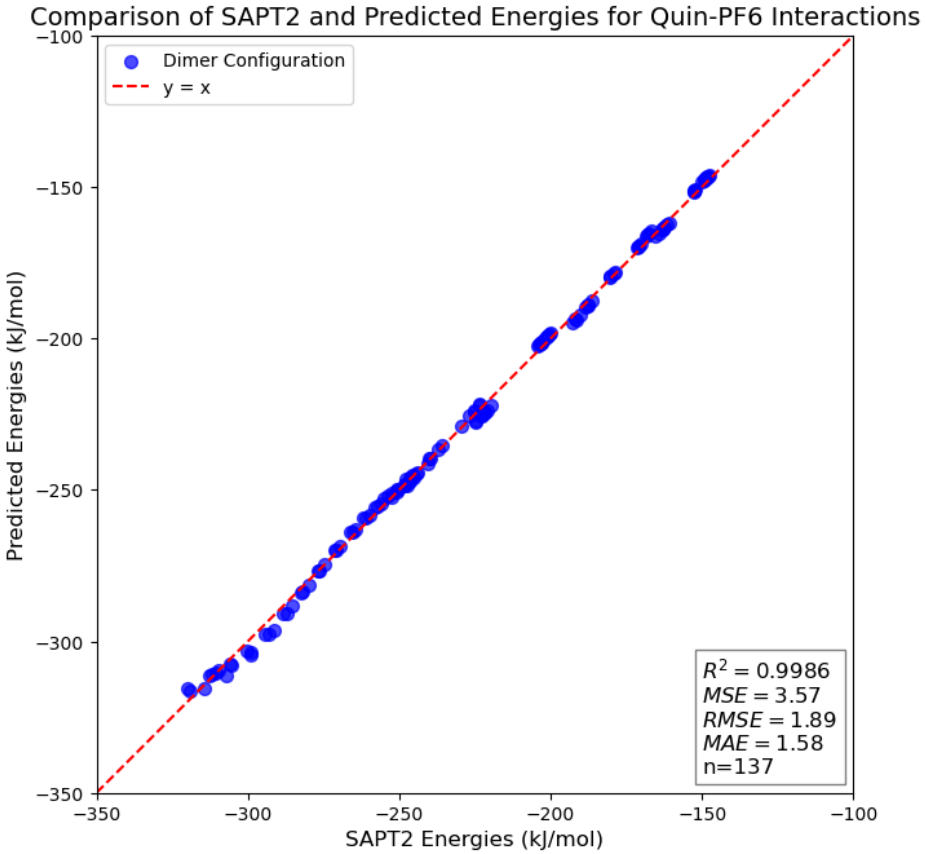


Figure 4.17: Parity plot comparing SAPT2 and predicted intermolecular energies (Buckingham potential + electrostatics), including metrics: $R^2 = 0.9986$, MSE = 3.57, RMSE = 1.89, MAE = 1.58, and $n = 137$.

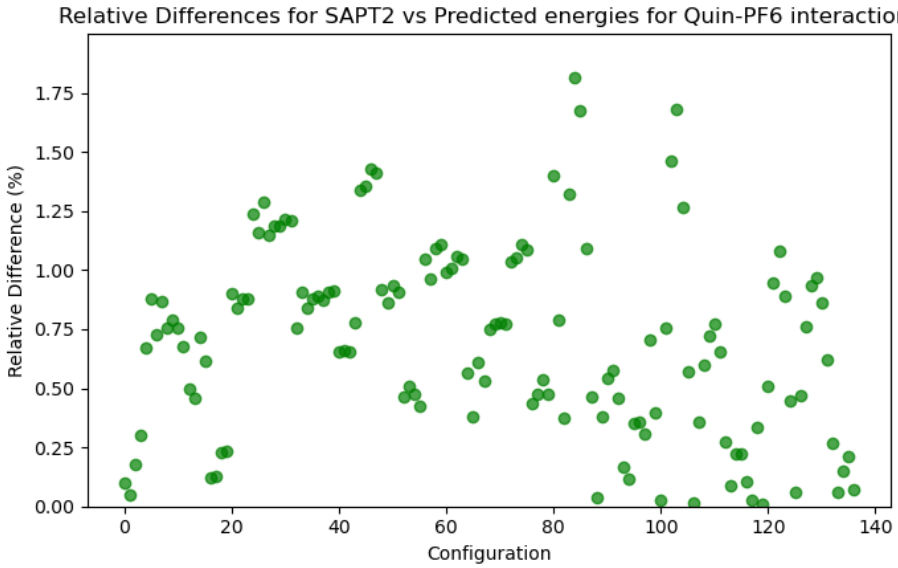


Figure 4.18: Relative error between predicted intermolecular energies (Buckingham potential + electrostatics) and SAPT2 energies for Quin-PF6 interactions.

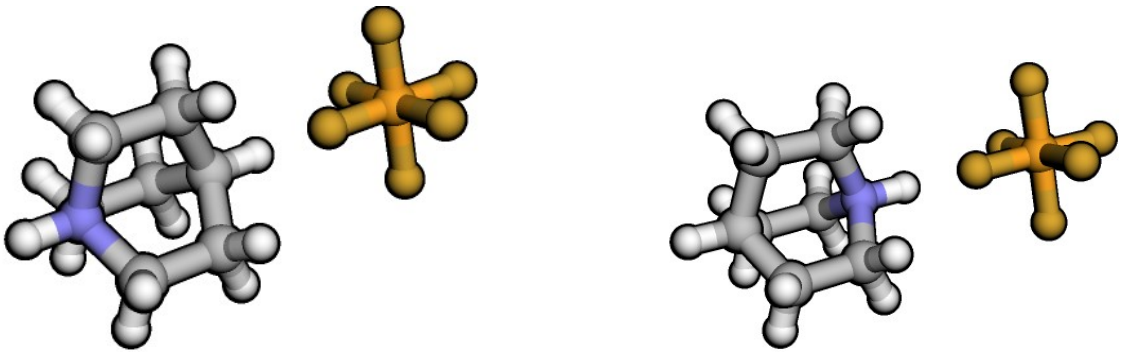
The comparison between SAPT2 and the predicted intermolecular energies shows good agreement, as illustrated by the scatter plot and parity plot. The R^2 value of 0.9986 confirms the accuracy of the fitting, with all relative errors below 2% across all configurations. This also ensures that these A, B, and C constant values can be used to model the system's intermolecular interactions in molecular dynamics simulations with high precision.

Considering all the information gathered and the analyses performed, it becomes evident that the initial fitting attempts for PF6-PF6 interactions highlighted that the molecular orientations had minimal influence on the interaction energies due to the high symmetry of the PF6 molecule. Instead, the distance between the molecules played a much larger role in determining the intermolecular forces. This is expected for a highly symmetric molecule like PF6, where the interaction energies are more sensitive to the separation between the molecules rather than their relative orientation. Despite this, early optimizations yielded suboptimal results, as reflected in the low R^2 value (0.7597) and significant relative errors. This suggests that even small deviations in parameterization can affect energy predictions, likely due to the subtle intermolecular forces present in this highly symmetric system. The subsequent use of the Nelder-Mead algorithm dramatically improved the fit, achieving an R^2 value of 0.9865 and relative errors below 2.5%. This confirms the robustness of the optimization, with the scatter and parity plots showing that the adjusted Buckingham parameters better capture the interaction energies for the PF6-PF6 system.

For QUIN-QUIN interactions, although the molecule is slightly less symmetric and its charge distribution is somewhat less uniform compared to PF6, the initial results were already promising. This is likely due to the fact that the initial parameters were previously optimized by Williams (2001). The initial R^2 value of 0.9724 was already quite high, and further optimization improved it to 0.9852, with relative errors dropping below 1.8%. This demonstrates the robustness of the initial parameters and the success of the refinement for this system.

In contrast, the QUIN-PF6 system presented more significant challenges. The largest errors occurred when the molecules rotated and approached each other closely, with the maximum error in the optimised model observed when the PF6 molecule was perfectly aligned and in close proximity

to the nitrogen-hydrogen bond in the QUIN molecule. Initially, before optimisation, the highest error (52.70%) occurred when a C-H bond was near a fluorine atom in PF6. After optimisation, although the error decreased significantly, the largest error shifted to the N-H bond, likely due to fewer configurations involving N-H interactions, slightly limiting the model's ability to fully adjust these short-range interactions, and underscoring the importance of molecular distance on the energy.



(a) (b)

Figure 4.19: (a) Configuration before the optimization of all the parameters, showing the largest relative error of 52.70% (Quin-PF6). (b) Configuration after the optimization, showing the largest relative error of 1.81% (Quin-PF6).

These results highlight that the distance between molecules is not only critical but also highly influenced by rotational effects. This sensitivity to orientation is particularly evident in the QUIN-PF6 interaction, where rotational effects have a more pronounced impact compared to the symmetric PF6-PF6 or QUIN-QUIN cases. The initial combined parameters, derived from individual PF6 and QUIN fits, were insufficient, yielding an R^2 value of 0.8896. However, after simultaneous optimisation, the fit improved significantly, with an R^2 value of 0.9986 and relative errors below 2%. The application of a penalty function during the optimisation process was crucial to maintaining physical consistency while allowing for slight adjustments to the parameters.

These findings are consistent with the nature of the system—a highly symmetric ionic crystal with molecules that have considerable rotational freedom. The strong dependence on orientation for QUIN-PF6 interactions likely arises from the close-range electrostatic interactions between the nitrogen-hydrogen bond and PF6, where alignment plays a critical role. The fitting now provides an accurate representation of these complex interactions, ensuring that the Buckingham parameters can be confidently applied in molecular dynamics simulations to model both symmetric and asymmetric interactions.

4.1.4 System simulation and phase transitions identification

To obtain the system trajectories and analyze whether the behavior matched the expected results, a fully functional forcefield was required. The intermolecular interactions were accurately defined using the potentials and constants, as described in the previous section. For the remaining parameters in the forcefield, such as bonds, angles, and dihedrals, the OPLS-AA forcefield was employed. This approach resulted in a complete and functional forcefield, enabling the molecular dynamics simulations to run in OpenMM, while also ensuring that the intermolecular interactions—which are likely to have a strong influence in the barocaloric behavior of the system—were well-defined.

The simulations were conducted successfully, with no issues related to the stability of the forcefield or the integrity of the software. Throughout the simulation range from 200K to 500K with 25000

time steps each simulation, where the phase transition behavior is expected, key indicators such as energy conservation, temperature, and pressure remained within expected ranges, confirming the robustness of the system. Additionally, no anomalies, such as bond stretching or system instability, were observed, ensuring the reliability of the computational setup. These preliminary results confirm that the simulation proceeded as intended, providing a solid foundation for further structural and dynamic analysis.

After running stable simulations with the optimized forcefield, the resulting trajectories were saved for further use in developing a neural network-based forcefield. Additionally, several diagrams illustrating the H-N bond orientations were generated to provide insight into the dynamics of the system, using these bonds in Quin molecules as indicators. The resulting diagrams are as follows:

200K

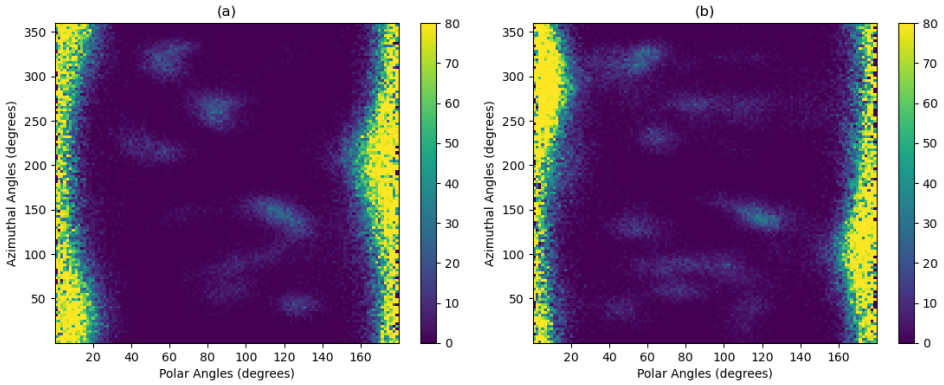


Figure 4.20: Heat maps showing the most recurrent positions of the H-N bond in the Quin molecule, indicating molecular movements. Different friction coefficients were used in the simulations: (a) 200K with a friction coefficient of 2 ps^{-1}, and (b) 200K with a friction coefficient of 3.5 ps^{-1}.

300K

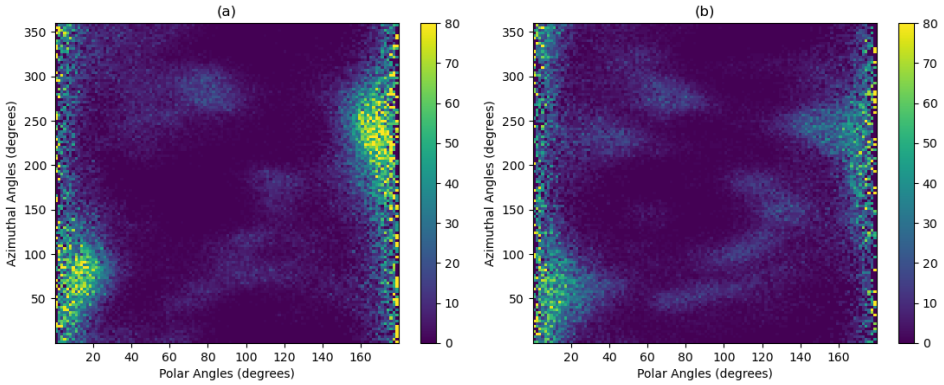


Figure 4.21: Heat maps showing the most recurrent positions of the H-N bond in the Quin molecule, indicating molecular movements. Different friction coefficients were used in the simulations: (a) 300K with a friction coefficient of 1 ps^{-1}, and (b) 300K with a friction coefficient of 4 ps^{-1}.

350K

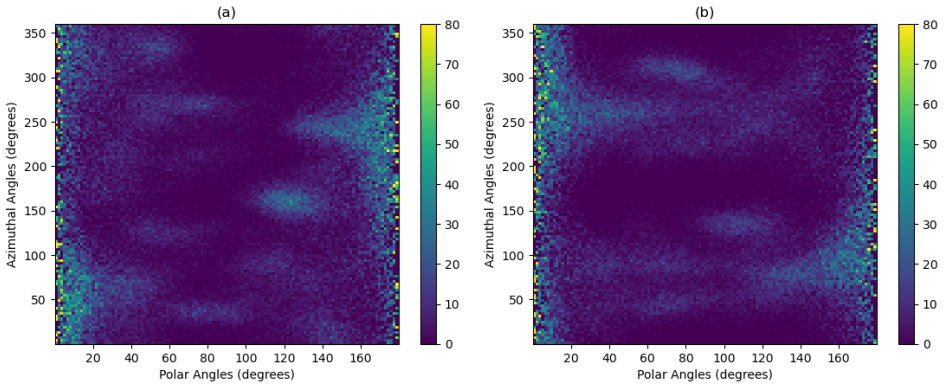


Figure 4.22: Heat maps showing the most recurrent positions of the H-N bond in the Quin molecule, indicating molecular movements. Different friction coefficients were used in the simulations: (a) 350K with a friction coefficient of 2 ps^{-1}, and (b) 350K with a friction coefficient of 3.5 ps^{-1}.

400K

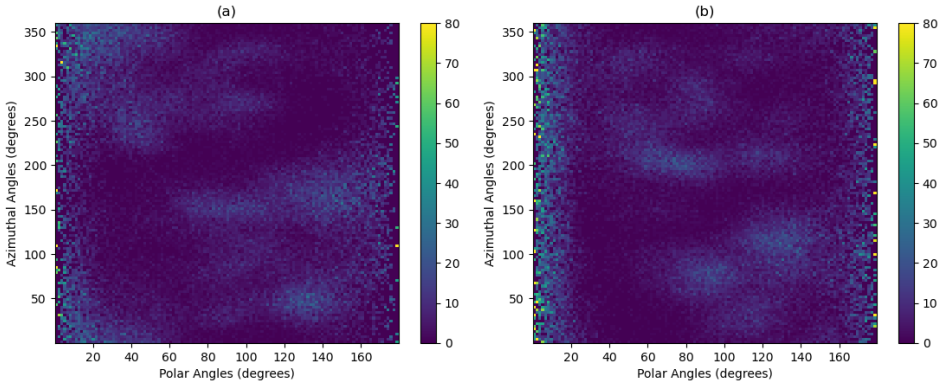


Figure 4.23: Heat maps showing the most recurrent positions of the H-N bond in the Quin molecule, indicating molecular movements. Different friction coefficients were used in the simulations: (a) 400K with a friction coefficient of 2.5 ps^{-1}, and (b) 400K with a friction coefficient of 3 ps^{-1}.

500K

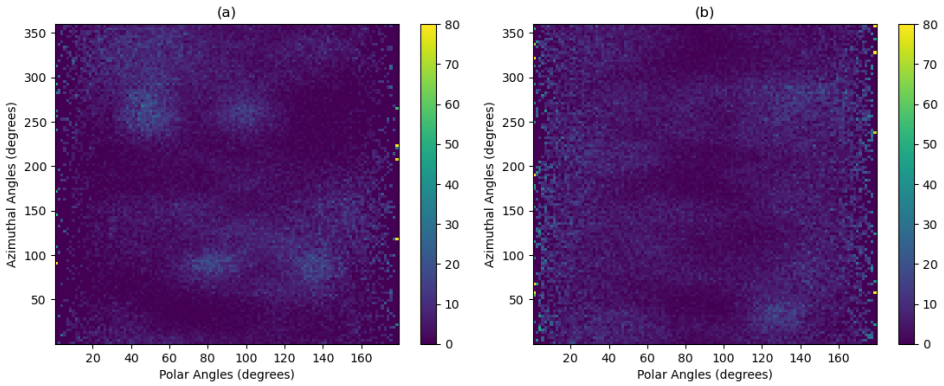


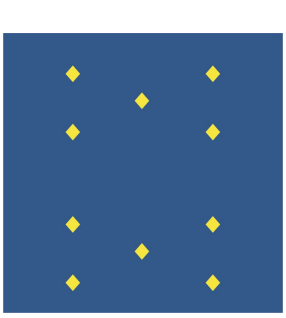
Figure 4.24: Heat maps showing the most recurrent positions of the H-N bond in the Quin molecule, indicating molecular movements. Different friction coefficients were used in the simulations: (a) 500K with a friction coefficient of 2.5 ps^{-1}, and (b) 500K with a friction coefficient of 3 ps^{-1}.

These heat maps reflect the physical movement of the Quin molecules and can be used to understand and verify the dynamics of the system under different temperature conditions.

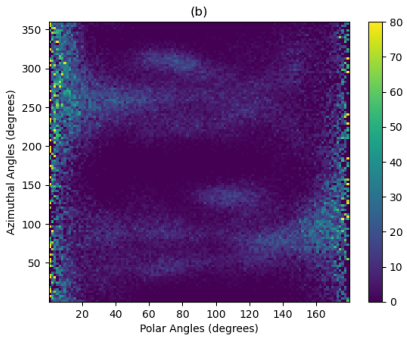
Molecular dynamics simulations, employing the OPLS-AA forcefield for bonds, angles, and dihedrals, and augmented by an intermolecular interactions model integrating electrostatics from CAMCASPT and parameters of the Buckingham potential refined in the previous sections, consistently demonstrated robust stability and outcomes that were in alignment with the predetermined ensembles throughout each simulation stage. To further substantiate that the system's behavior is congruent with its actual physical properties, data were utilized from X-ray crystallography on a Quin-PF6 sample disclosed a specific orientation pattern of molecular structures during high-entropy phases and proximate to phase transitions, with the N-H bond serving as a prominent indicator.

As the system's energy escalated, providing increased rotational freedom to the molecule, the orientation pattern exhibited greater dispersion. This dispersion is attributable to the fact that while an increase in energy facilitates the alignment of the Quin molecule towards its preferred orientation points, an excess of energy induces erratic movements, deviating orientations from the established pattern. The fidelity of this pattern replication in simulations is corroborated by the attached images, which depict the orientation of N-H bonds in high-entropy regions near phase transitions, and a heatmap of N-H bond orientations at temperatures approaching these transitions.

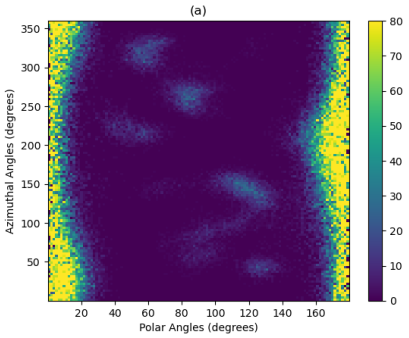
Additionally, an image capturing a very low entropy state illustrates the initial orientations of the N-H bonds, documenting the system's evolution between different molecular orientations in the Quin molecule. These findings underscore the forcefield's capacity to accurately simulate real-world behaviors, substantiating its applicability across diverse Quin-PF6 related applications.



(a) Orientation pattern of the N-H bond in high-entropy regions.



(b) N-H bond orientations at 350K, near the phase transition showing a similar pattern.



(c) Initial orientations of the N-H bond at a low entropy state at 200K.

Figure 4.25: Observed orientation patterns of N-H bonds in high-entropy phases near phase transitions, compared across different temperature conditions within the Quin-PF6 system.

The expected behaviors under high-temperature conditions, which correspond to states of increased entropy, are clearly demonstrated in the simulations. These conditions lead to molecules displaying a variety of orientations and some degree of disorder, yet they maintain their structural arrangement. This is primarily due to in-place rotations. The forthcoming images, taken from the resulting trajectories of the simulations, explicitly illustrate these phenomena, showing how the simulations at high temperatures capture the dynamic yet structurally intact behaviour of the molecules.

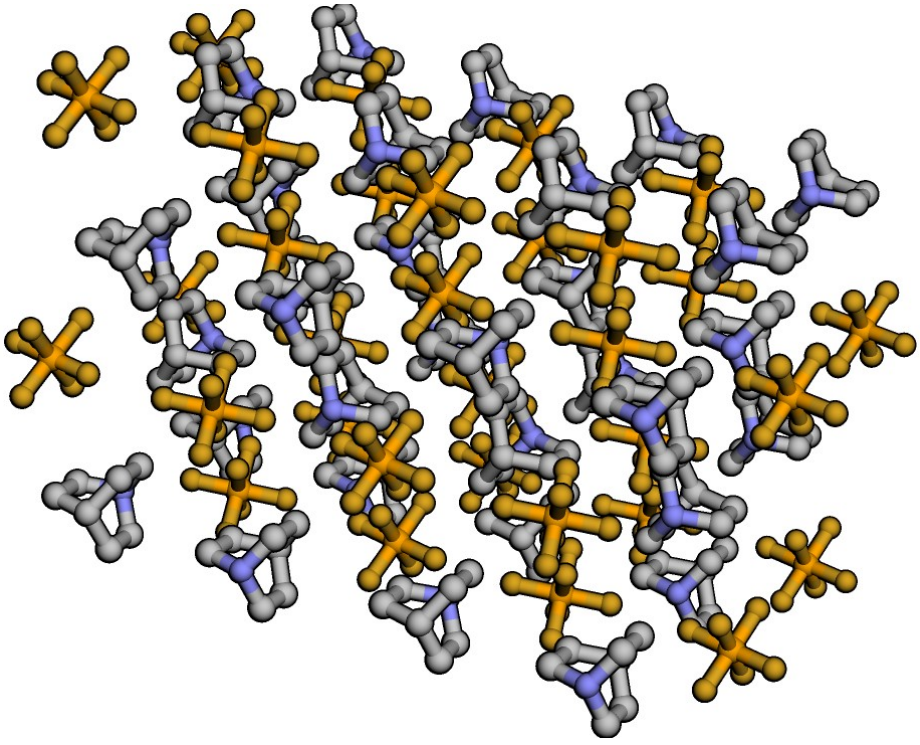


Figure 4.26: Structure of Quin-PF6 showing clear orientations and ordered structure.

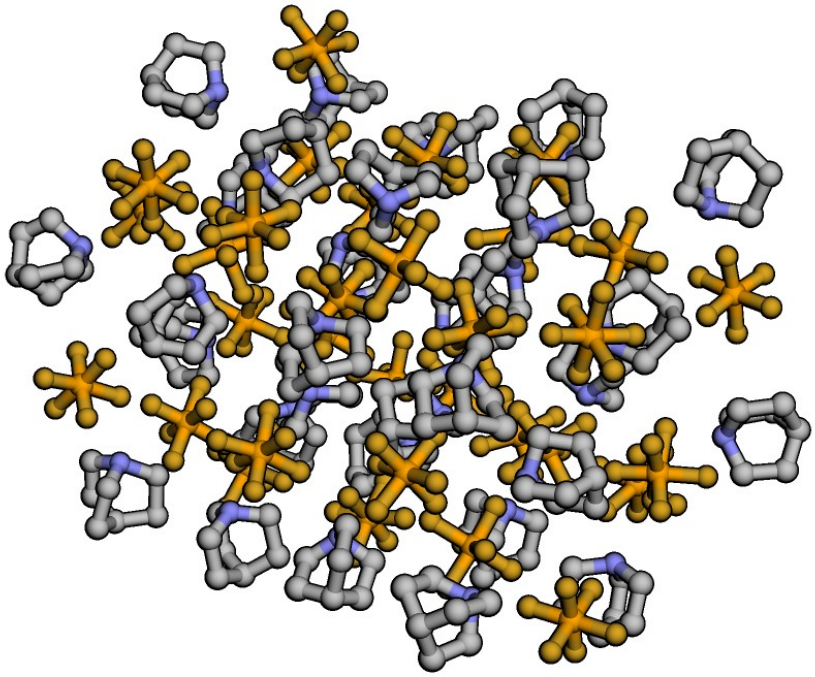


Figure 4.27: High entropy phase at 500K, depicting various orientations of Quin molecules while maintaining a organized structural arrangement.

These images illustrate the expected behavior of the system in simulations using the fitted forcefield. Additionally, with the appropriate visualization software, the model can be utilized to analyze and understand various phenomena in the Quin-PF6 system, confirming both its utility and similarity with real life physics of the system.

4.2 Neural Network Fitting for Energy Predictions

Neural networks are widely recognized for their ability to uncover relationships in complex systems, especially where functional dependencies are not straightforward or when the system involves discontinuous functions. In the context of intermolecular interactions and potential energies, it is often challenging to determine the most appropriate potential function or even to fit such a function using traditional methods. Neural network-based forcefields offer an alternative approach by smoothing predictions, enhancing accuracy, and extending the predictive capabilities of classical forcefields, particularly through extrapolation due to their descriptor based nature.

For the current study, SOAP (Smooth Overlap of Atomic Positions) descriptors were utilized to represent the system's configurations. The dataset consisted of 12,510 instances, each derived from snapshots of the system's trajectory files. Specifically, each trajectory file contained 1,251 snapshots corresponding to different temperature and friction combinations, with two friction coefficients associated with each temperature for a grat total of the 12510 observations mentioned. Additionally, each instance was characterized by 3,780 variables (SOAP descriptors), which at the same time underscores the high dimensionality of the problem. The goal was to capture the energy patterns encoded within the classical forcefield parameters through neural network models.

The final dataset dimensions were (12510, 3781), where the last column represents the target variable, which corresponds to the potential energy, this column was evaluated as a target to get a regression model. Different neural network architectures were explored to determine the most suitable model for fitting these energy patterns.

The primary approach decided upon involved generating validation graphs and using three main performance metrics—R^2, MAE, and RMSE. This strategy was chosen to maintain a targeted assessment of the model's generalization capabilities. By exclusively using the validation set, the aim was to reduce bias that could result from the inclusion of training data, thus offering a more precise understanding of the model's performance on new, unseen data.

Focusing on R^2 allowed for an assessment of the proportion of variance explained by the model, which is crucial for understanding its predictive power. Meanwhile, MAE and RMSE offered complementary perspectives on error magnitude, with MAE providing a straightforward measure of average error and RMSE emphasizing larger errors. This combination of metrics provides a comprehensive view of the model's performance without overwhelming complexity.

Moreover, limiting the evaluation to these specific metrics ensured that the analysis remained interpretable, facilitating the identification of areas for improvement.

4.2.1 Feedforward Neural Network

As a starting point, a basic feedforward neural network (FFNN) was tested. Despite its simplicity, the FFNN is often useful in many scenarios, which makes it a logical initial approach. The network architecture included three hidden layers, with up to 1,024 neurons per layer at maximum capacity. The feedforward network serves as a benchmark for more complex models, allowing the evaluation of the potential of neural networks to make good predictions in this type of data.

The performance of the FFNN was the following:

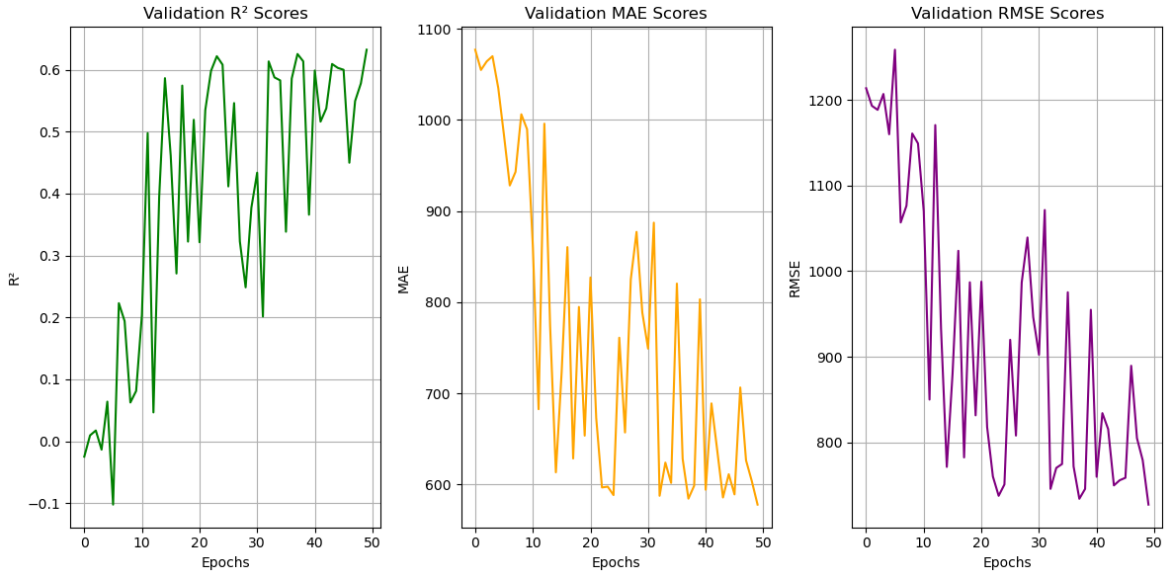


Figure 4.28: Validation metrics for the basic feedforward neural network.

The results indicate an erratic behavior, with significant peaks and valleys observed in the graphs. Despite this volatility, the **R^2 score** shows a slight upward trend, while both **MAE** and **RMSE** exhibit a downward trajectory. However, after 50 epochs, the improvements are minimal, with the R^2 reaching a maximum of only 0.644 before stagnating. This behavior could be attributed to insufficient network complexity to fully capture the intricacies of the energy landscape or an inadequate learning rate that fails to stabilize convergence.

4.2.2 Feedforward Neural Network with more layers

To test the effect of a deeper architecture, a feedforward neural network with five hidden layers, each containing 512 neurons, was constructed. The hypothesis was that increasing the depth of the network could help in capturing more intricate energy patterns that may not have been fully represented in the three-layer architecture.

The performance of the FFNN was the following:

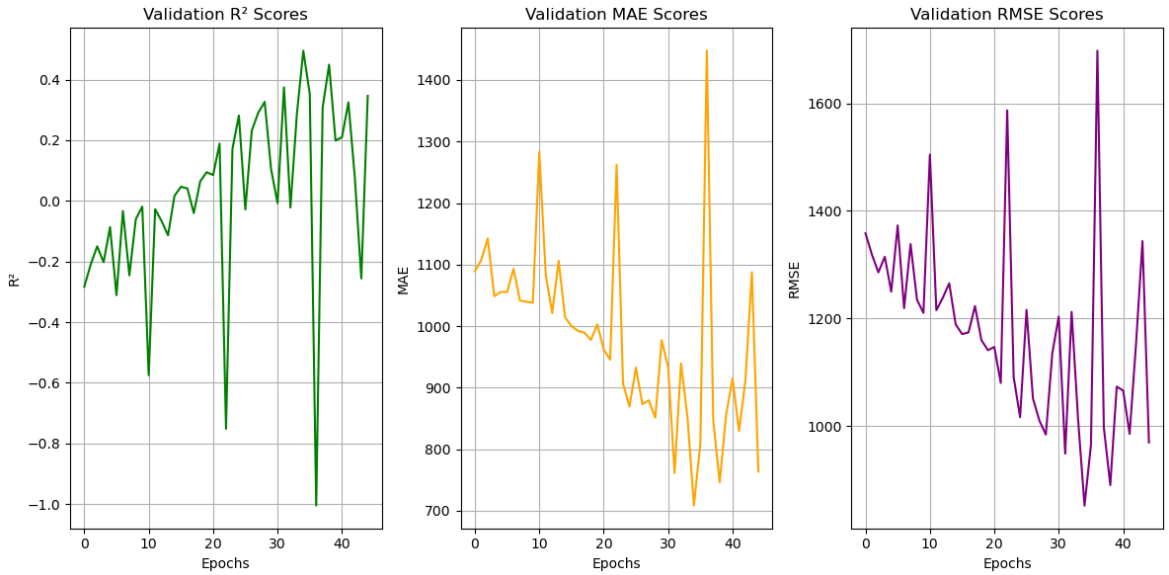


Figure 4.29: Validation metrics for the feedforward neural network with more layers.

The analysis reveals a slightly more stable behavior compared to the previous model; however, when peaks occur, they are significantly larger than the troughs. This suggests that while the model has improved in stability, it is still susceptible to overfitting, as indicated by the R^2 score of 0.526. The larger peaks might reflect the model's attempt to fit outliers or noise in the training data, demonstrating a potential misalignment with the underlying data distribution.

4.2.3 Feedforward Neural Network with More Neurons and More Layers

In order to test the effect of a deeper architecture, a feedforward neural network with six hidden layers, each containing 1024 neurons, was constructed.

The performance of the FFNN was the following:

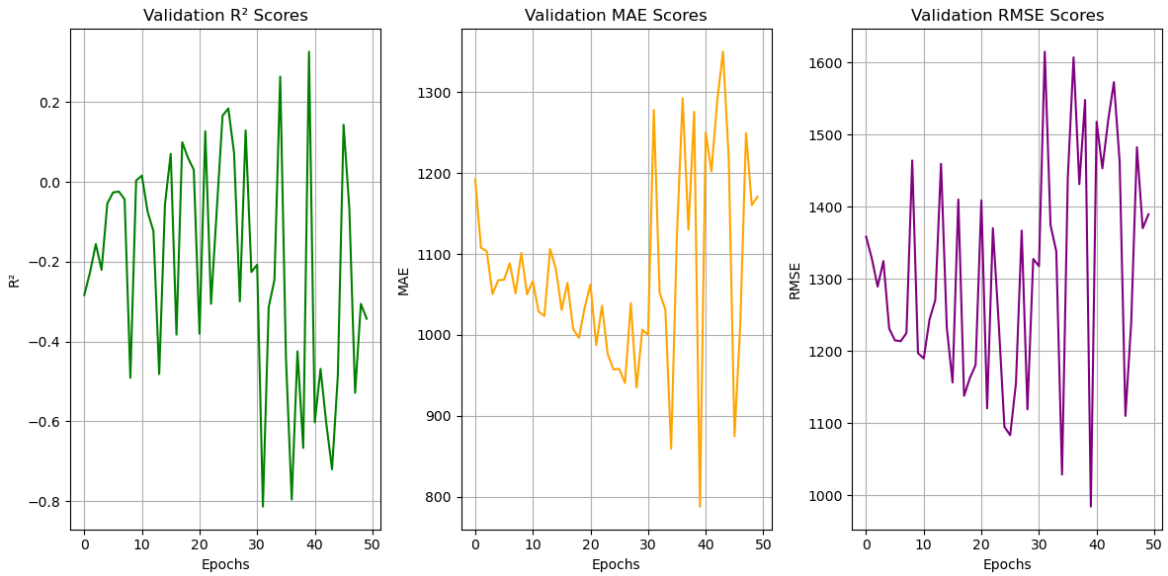


Figure 4.30: Validation metrics for the feedforward neural network with additional layers.

The results exhibit a pattern resembling a cage, where the peaks of increases and decreases are notably similar. Although there is a slight upward trend, it diminishes rapidly, suggesting that the model struggles to effectively capture the underlying energy patterns. This could indicate that the additional layers are not contributing positively to the model's learning, perhaps due to saturation or gradient issues often encountered in deeper networks.

4.2.4 Feedforward Neural Network with AdamW Optimizer

To further refine the performance, the AdamW optimizer was introduced in the FFNN architecture with three hidden layers and 512 neurons. AdamW is known for its efficiency in dealing with high-dimensional spaces and noisy gradients, making it an appropriate choice for the current problem.

The performance of the FFNN with adamW optimizer is the following:

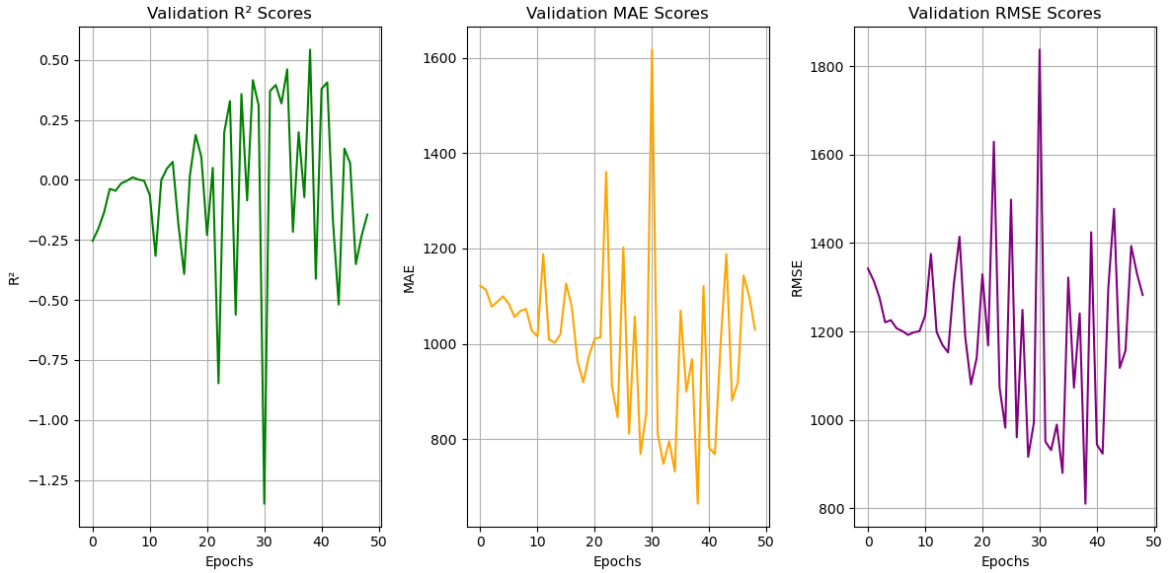


Figure 4.31: Validation metrics for the feedforward neural network with AdamW optimizer.

The graph shows an erratic movement with alternating peaks and valleys, resembling a mini-grid pattern. Despite some fluctuations, there is only a very slight upward trend, suggesting that the model is not converging effectively. This could be due to the dropout rate introduced (0.003), which might be too low to significantly impact regularization, or it might indicate that the optimizer alone is insufficient to stabilize training.

4.2.5 Feedforward Neural Network with AdamW Optimizer with More Neurons

To further refine the performance, the AdamW optimizer was introduced in the FFNN architecture with more neurons (1024).

The performance of the FFNN with adamW(1024) was the following :

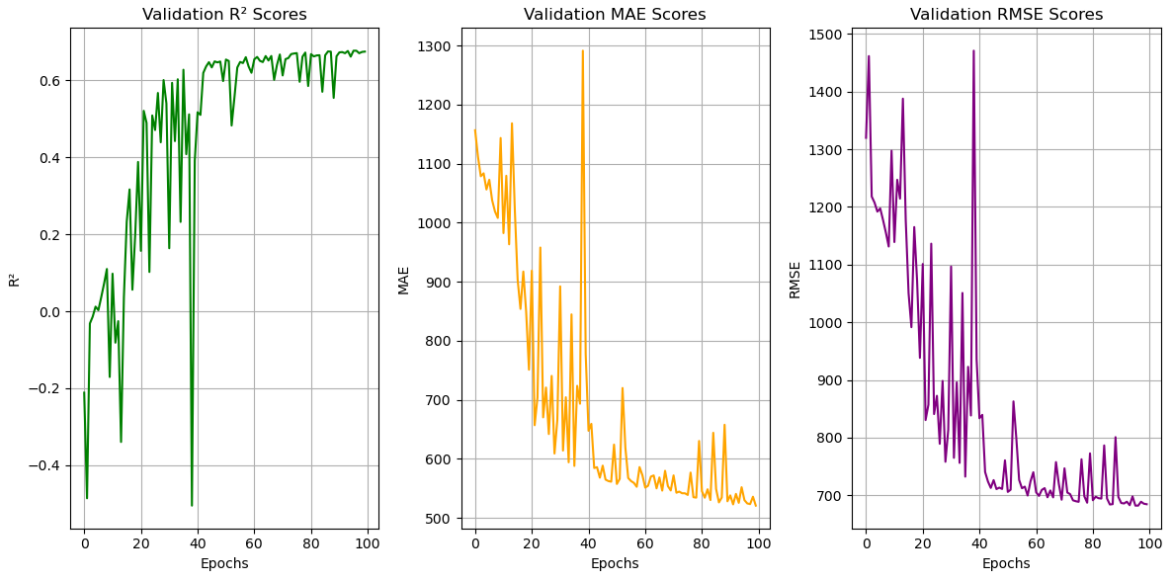


Figure 4.32: Validation metrics for the FFNN with AdamW optimizer and increased neurons.

This configuration showed a stable performance around an R^2 score of approximately 0.7. While stability is a positive aspect, the lack of significant changes suggests that the model was stuck and cannot improve further. The observed similarity between the R^2 values and the corresponding MAE and RMSE indicates a consistent prediction performance but too low to be useful.

4.2.6 Convolutional Neural Network

A convolutional neural network (CNN) was implemented to investigate whether spatial patterns in the SOAP descriptors could be leveraged to improve model accuracy.

The performance of the CNN was the following:

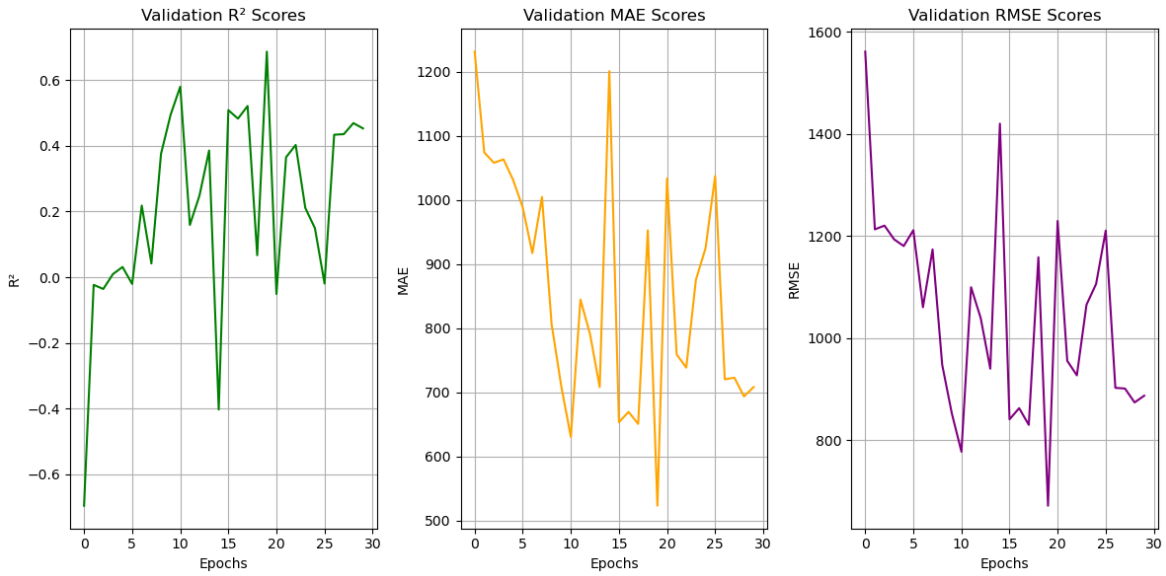


Figure 4.33: Validation metrics for the convolutional neural network.

The CNN demonstrates slower movement with less rapid fluctuations, yet it maintains high differences between minimum and maximum values. The significant variability suggests that the model struggles to find a definitive trend in either direction, indicating challenges in effectively capturing the spatial relationships encoded in the SOAP descriptors. The ability of CNNs to leverage local features may still provide some advantage, but the overall inconsistency suggests further tuning is needed.

4.2.7 Long Short-Term Memory (LSTM) Network

Finally, an LSTM network was implemented to explore whether temporal dependencies between the sequential data of molecular dynamics could be utilized.

The performance of the LSTM measured was the following:

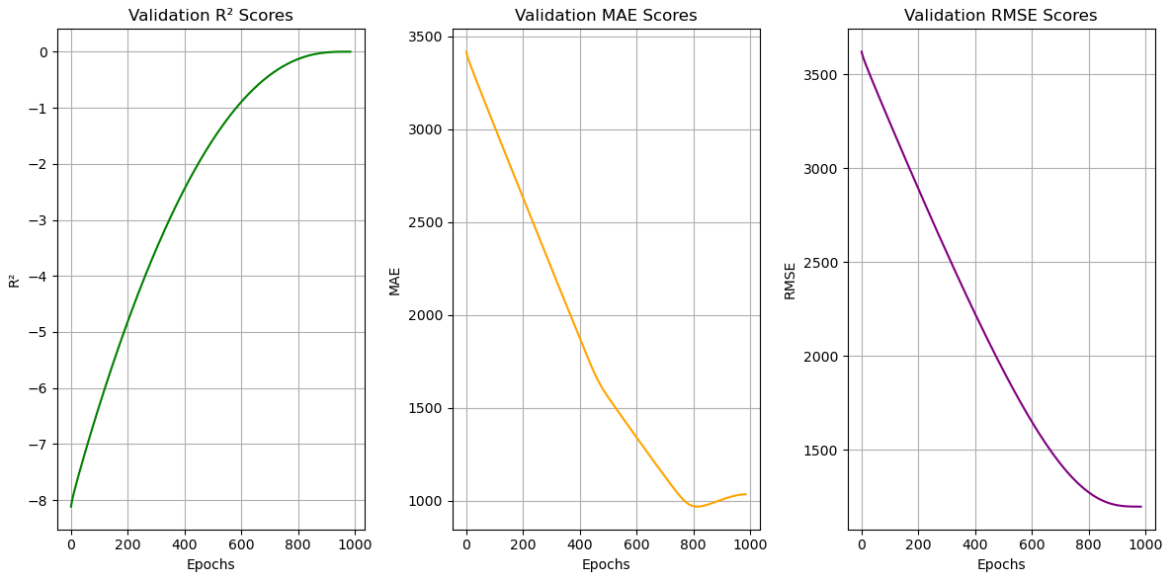


Figure 4.34: Validation metrics for the LSTM network.

The LSTM network exhibited a smooth trajectory, with gradual changes in performance metrics.

Starting from a negative R^2 score, it ended at approximately zero, raising questions about its overall effectiveness. The transition from around -8 to 0 suggests a learning process that is too slow or potentially misaligned with the data. The MAE and RMSE metrics followed a similarly smooth declining trend, indicating that the model is making minor adjustments, but it ultimately fails to capture the underlying dynamics of the system.

5 Conclusions and Further Work

5.1 Conclusions

From the results presented in this thesis the following conclusion can be draw:

1. **Ab initio calculations and forcefield fitting**:

   The *ab initio* calculations performed on the Quin-PF6 system were essential for obtaining precise parameters, enabling the proper fitting of the Buckingham potential. The efforts in forcefield parametrization and fitting were successful in capturing the key intermolecular interactions, as evidenced by the excellent correlation between the energies calculated through SAPT2 and the energies predicted by the optimized forcefield. The results demonstrate that the forcefield is robust and well-calibrated to describe non-bonded interactions, which is crucial for accurately representing the energy behavior of the Quin-PF6 system and allowing to show barocaloric effects.

2. **Fitting of intermolecular interactions**:

   The optimisation of PF6-PF6 and QUIN-QUIN interactions was successful after several adjustments, with correlation coefficients R^2 above 0.98 and relative errors below 2%. However, the optimisation of the QUIN-PF6 system was more challenging due to the increased complexity of interactions between the two molecules. Although initial attempts showed significant errors, the final optimisation using the Nelder-Mead algorithm resulted in a satisfactory fit, with an R^2 of 0.9986 and relative errors below 2%, confirming that the parameters obtained are suitable for modelling the system and the algorithm present a reliable option for optimise forcefield although is not the most common used.

3. **Impact of molecular symmetry and orientation**: Observations during the forcefield parameter optimization showed that the symmetry of the system, particularly in PF6-PF6 interactions, minimizes the impact of molecular orientations on interaction energies. In contrast, for the QUIN-PF6 system, the orientation and proximity between molecules played a crucial role in the accuracy of energy predictions, highlighting the importance of considering both distance and orientation in asymmetric systems.

4. **Simulations and comparison with experimental data**:

   Molecular dynamics simulations using the optimized forcefield successfully replicated the physical behaviors observed in real-world experimental data. The simulations effectively identified phase transitions and barocaloric behavior, aligning with experimental observations obtained from techniques like quasi-elastic electron scattering. The stability and consistency of the system over the temperature range, as well as the agreement with theoretical predictions, confirm the validity of the model in simulating real systems.

5. **Descriptors and neural networks**:

   Despite considerable efforts to use SOAP descriptors for training a neural network to accurately predict the energy behaviour of the Quin-PF6 system, the results were limited. The correlation in neural network training did not exceed 0.70 in any of the tested cases, even after exploring various architectures, including convolutional neural networks (CNNs), LSTM networks, and traditional feedforward networks, as well as multiple optimizers, such as Adam and AdamW. This suggests that the SOAP descriptors may not sufficiently capture all relevant features of the system. Furthermore, increasing the number of data points doesn't seem to make a difference in the performance, as the final dataset already contained 12,510 instances, yet no significant correlation values were observed. Thus, more advanced or alternative descriptors may be necessary to enhance the model's performance.

5.2 Further Work

Although the results of this project have successfully demonstrated the robustness of the optimized forcefield for the Quin-PF6 system, several areas remain open for future exploration. One critical direction involves improving the representation of molecular interactions using advanced descriptors. As the current SOAP descriptors were not sufficient to capture the full complexity of the intermolecular interactions in the Quin-PF6 system, further research should focus on developing or adopting alternative descriptors. Methods such as graph-based or machine learning-enhanced descriptors could be explored to better characterize the intricate intermolecular forces and improve the neural network's ability to predict energy behavior with higher accuracy.

Despite employing a variety of architectures and optimizers, the neural network models created in this project showed limited performance. Future studies might focus on developing more advanced neural network models or hybrid methods that combine machine learning with traditional physical models. Enhancing the dataset by incorporating simulation data from extreme temperatures could also improve predictive capabilities. Additionally, expanding the training set with more ab initio data might offer a stronger foundation for precise neural network predictions. Lastly, integrating the neural network-based forcefield into molecular dynamics simulation software and assessing its real-time performance against classical forcefields could validate its practical effectiveness for simulating barocaloric materials.

Future research on the Quin-PF6 system should aim to better elucidate the complex interactions between quinuclidinium and hexafluorophosphate ions, with an emphasis on their barocaloric properties. Although the current forcefield captures key intermolecular interactions, further refinements are required to optimize QUIN-PF6 interactions. Prospective studies should employ a combination of approaches and draw comparisons with similar systems to gain a more detailed understanding of the characteristics of these materials and the accurate influence on the barocaloric effect due to their ionic nature.

Bibliography

[1] Sten Nilsson and D. Pitt. *Protecting the Atmosphere: the Climate Change Convention and its context*. International Environmental Governance Set. Taylor and Francis. OCLC: 863823224.

[2] Leslie Daryl Danny Harvey. *Global warming the hard science*. Routledge. OCLC: 1289809232.

[3] Adrian Mota Babiloni, Angelo Maiorino, and Ciro Aprea. *Refrigeration Systems and Applications 2019*. Online access: OAPEN DOAB Directory of Open Access Books. MDPI AG 2019. OCLC: 1229184598.

[4] David Munoz-Rojas and Xavier Moya. *Materials for Sustainable Energy Applications*. Jenny Stanford Publishing, 0 edition.

[5] Carlos Escorihuela–Sayalero, Luis Carlos Pardo, Michela Romanini, Nicolas Obrecht, Sophie Loehlé, Pol Lloveras, Josep-Lluís Tamarit, and Claudio Cazorla. Prediction and understanding of barocaloric effects in orientationally disordered materials from molecular dynamics simulations. 10(1):1–10.

[6] Qingyong Ren, Ji Qi, Dehong Yu, Zhe Zhang, Ruiqi Song, Wenli Song, Bao Yuan, Tianhao Wang, Weijun Ren, Zhidong Zhang, Xin Tong, and Bing Li. Ultrasensitive barocaloric material for room-temperature solid-state refrigeration. 13(1):2293.

[7] Pol Lloveras and Josep-Lluís Tamarit. Advances and obstacles in pressure-driven solid-state cooling: A review of barocaloric materials. 8(1):3–15.

[8] X. Moya and N. D. Mathur. Caloric materials for cooling and heating. 370(6518):797–803.

[9] Drew Lilley and Ravi Prasher. Ionocaloric refrigeration cycle. 378(6626):1344–1348.

[10] M T Dove, D Fincham, and R E Hubbard. Dynamics of orientationally disordered crystals. 4(2):79–81.

[11] Haijin Zhu, Douglas R. MacFarlane, Jennifer M. Pringle, and Maria Forsyth. Organic ionic plastic crystals as solid-state electrolytes. 1(1):126–140.

[12] Bing Li, Yukinobu Kawakita, Seiko Ohira-Kawamura, Takeshi Sugahara, Hui Wang, Jingfan Wang, Yanna Chen, Saori I. Kawaguchi, Shogo Kawaguchi, Koji Ohara, Kuo Li, Dehong Yu, Richard Mole, Takanori Hattori, Tatsuya Kikuchi, Shin-ichiro Yano, Zhao Zhang, Zhe Zhang, Weijun Ren, Shangchao Lin, Osami Sakata, Kenji Nakajima, and Zhidong Zhang. Colossal barocaloric effects in plastic crystals. 567(7749):506–510.

[13] David Boldrin. Fantastic barocalorics and where to find them. 118(17):170502.

[14] Luca Cirillo, Adriana Greco, and Claudia Masselli. Cooling through barocaloric effect: A review of the state of the art up to 2022. 33:101380.

[15] A. M. G. Carvalho, W. Imamura, E. O. Usuda, and N. M. Bom. Giant room-temperature barocaloric effects in PDMS rubber at low pressures. 99:212–221.

[16] Araceli Aznar, Pol Lloveras, María Barrio, Philippe Negrier, Antoni Planes, Lluís Mañosa, Neil D. Mathur, Xavier Moya, and Josep-Lluís Tamarit. Reversible and irreversible colossal barocaloric effects in plastic crystals. 8(2):639–647.

[17] M. V. Gorev, E. V. Bogdanov, I. N. Flerov, A. G. Kocharova, and N. M. Laptash. Investigation of thermal expansion, phase diagrams, and barocaloric effect in the (NH4)2wo2f4 and (NH4)2moo2f4 oxyfluorides. 52(1):167–175.

[18] Wei-Jian Xu, Peng-Fei Li, Yuan-Yuan Tang, Wei-Xiong Zhang, Ren-Gen Xiong, and Xiao-Ming Chen. A molecular perovskite with switchable coordination bonds for high-temperature multiaxial ferroelectrics. 139(18):6369–6375.

[19] Steven P. Vallone, Anthony N. Tantillo, António M. Dos Santos, Jamie J. Molaison, Rafal Kulmaczewski, Antonin Chapoy, Pezhman Ahmadi, Malcolm A. Halcrow, and Karl G. Sandeman. Giant barocaloric effect at the spin crossover transition of a molecular crystal. 31(23):1807334.

[20] Ya. Grosu, V. Eroshenko, J. M. Nedelec, and J. P. E. Grolier. A new working mode for molecular springs: water intrusion induced by cooling and associated isobaric heat capacity change of a {ZIF-8 + water} system. 17(3):1572–1574.

[21] C. Nicolini, R. Ravindra, B. Ludolph, and R. Winter. Characterization of the temperature- and pressure-induced inverse and reentrant transition of the minimum elastin-like polypeptide GVG(VPGVG) by DSC, PPC, CD, and FT-IR spectroscopy. 86(3):1385–1392.

[22] Barocaloric effects in the solid-state: materials and methods.

[23] Susobhan Das, Amit Mondal, and C. Malla Reddy. Harnessing molecular rotations in plastic crystals: a holistic view for crystal engineering of adaptive soft materials. 49(24):8878–8896.

[24] *Computational soft matter: from synthetic polymers to proteins. lect: Lecture notes.* Number 23 in NIC series. NIC.

[25] M.A. González. Force fields and molecular dynamics simulations. 12:169–200.

[26] Abdelaziz Yasri, Laurent Chiche, Jacques Haiech, and Gerard Grassy. Rational choice of molecular dynamics simulation parameters through the use of the three-dimensional autocorrelation method: application to calmodulin flexibility study. 9(11):959–976.

[27] Jack Wildman, Peter Repiščák, Martin J. Paterson, and Ian Galbraith. General force-field parametrization scheme for molecular dynamics simulations of conjugated materials in solution. 12(8):3813–3824.

[28] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. de Groot, Helmut Grubmüller, and Alexander D. MacKerell. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. 14(1):71–73.

[29] Jay W. Ponder, Chuanjie Wu, Pengyu Ren, Vijay S. Pande, John D. Chodera, Michael J. Schnieders, Imran Haque, David L. Mobley, Daniel S. Lambrecht, Robert A. DiStasio, Martin Head-Gordon, Gary N. I. Clark, Margaret E. Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. 114(8):2549–2564.

[30] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. 118(45):11225–11236.

[31] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. 65(3):712–725.

[32] Nathan Schmid, Andreas P. Eichenberger, Alexandra Choutko, Sereina Riniker, Moritz Winger, Alan E. Mark, and Wilfred F. van Gunsteren. Definition and testing of the GROMOS force-field versions 54a7 and 54b7. 40(7):843–856.

[33] Martin H. Müser, Sergey V. Sukhomlinov, and Lars Pastewka. Interatomic potentials: achievements and challenges. 8(1):2093129.

[34] Frank H. Stillinger and Thomas A. Weber. Computer simulation of local order in condensed phases of silicon. 31(8):5262–5271.

[35] Zhennan Zhang, Yaxiong Chen, and Hong Zheng. A modified stillinger–weber potential-based hyperelastic constitutive model for nonlinear elasticity. 51(7):1542–1554.

[36] A. J. Stone. *The theory of intermolecular forces.* Oxford University Press, second edition published in paperback edition.

[37] D. C. Rapaport. *The art of molecular dynamics simulation.* Cambridge University Press.

[38] Philippe A. Bopp, Ewa Hawlicka, and Siegfried Fritzsche. The hitchhiker's guide to molecular dynamics. 4(1):2.

[39] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. 1-2:19–25.

[40] I. D. Brown. CIF (crystallographic information file): A standard for crystallographic data interchange. 101(3):341.

[41] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. 28(1):235.

[42] Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João Pglm Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. OpenMM 8: Molecular dynamics simulation with machine learning potentials. 128(1):109–116.

[43] B. R. Brooks, I. I. I. C. L. Brooks, Jr A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. 30(10):1545.

[44] David A. Case, Iii Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Jr Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The amber biomolecular simulation programs. 26(16):1668.

[45] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. 271:108171.

[46] John P. Lowe. *Quantum chemistry*. Academic Press, 2nd ed edition.

[47] Richard M. Martin. *Electronic structure: basic theory and practical methods*. Cambridge University Press.

[48] Alexandre M. Zagoskin. *Quantum mechanics: a complete introduction*. Teach yourself. John Murray Learning. OCLC: ocn927132599.

[49] Attila Szabo and Neil S. Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Dover Publications.

[50] Barton Zwiebach. *Mastering quantum mechanics: essentials, theory, and applications*. The MIT Press.

[51] Jos Thijssen. *Computational Physics*. Cambridge University Press, 2 edition.

[52] István Mayer. The hartree-fock method. In István Mayer, editor, *Simple Theorems, Proofs, and Derivations in Quantum Chemistry*, pages 165–225. Springer US.

[53] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. 140(4):A1133–A1138.

[54] Trent M. Parker, Lori A. Burns, Robert M. Parrish, Alden G. Ryno, and C. David Sherrill. Levels of symmetry adapted perturbation theory (SAPT). i. efficiency and performance for interaction energies. 140(9):094106.

[55] Bogumil Jeziorski, Robert Moszynski, and Krzysztof Szalewicz. Perturbation theory approach to intermolecular potential energy surfaces of van der waals complexes. 94(7):1887–1930.

[56] Reinaldo Padilha França, Ana Carolina Borges Monteiro, Rangel Arthur, and Yuzo Iano. Chapter 3 - an overview of deep learning in big data, image, and signal processing in the modern digital age. In Vincenzo Piuri, Sandeep Raj, Angelo Genovese, and Rajshree Srivastava, editors, *Trends in Deep Learning Methodologies*, Hybrid Computational Intelligence for Pattern Analysis, pages 63–87. Academic Press.

[57] Davide Anguita, Alessandro Ghio, Noemi Greco, Luca Oneto, and Sandro Ridella. Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

[58] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. 213:106504.

[59] Haosheng Xu, Dongheng Qian, and Jing Wang. Predicting many properties of crystals by a single deep learning model.

[60] Optimization for machine learning. OCLC: ocn701493361.

[61] Michael Walker. *Data cleaning and exploration with machine learning: get to grips with machine learning techniques to achieve sparkling-clean data quickly*. Packt Publishing. OCLC: 1341444522.

[62] Stefan Sandfeld. What you should know about data, math, and computing. In Stefan Sandfeld, editor, *Materials Data Science: Introduction to Data Mining, Machine Learning, and Data-Driven Predictions for Materials Science and Engineering*, pages 31–52. Springer International Publishing.

[63] Sufiyan Sajid, Abid Haleem, Shashi Bahl, Mohd Javaid, Tarun Goyal, and Manoj Mittal. Data science applications for predictive maintenance and materials science in context to industry 4.0. 45:4898–4905.

[64] Joe Reis and Matthew L. Housley. *Fundamentals of data engineering: plan and build robust data systems*. O'Reilly Media, first edition edition. OCLC: on1334138491.

[65] Biju Issac and Gajendra Pal Singh Raghava. EGPred: Prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches. 14(9):1756–1766.

[66] Anupam Sanghi and Jayant R. Haritsa. Synthetic data generation for enterprise DBMS. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3585–3588. ISSN: 2375-026X.

[67] Logan G. Wright, Tatsuhiro Onodera, Martin M. Stein, Tianyu Wang, Darren T. Schachter, Zoey Hu, and Peter L. McMahon. Deep physical neural networks trained with backpropagation. 601(7894):549–555.

[68] John D. Kelleher. *Deep learning*. The MIT press essential knowledge series. The MIT Press.

[69] Stefan Sandfeld. Introduction and general concepts of machine learning and data science. In Stefan Sandfeld, editor, *Materials Data Science: Introduction to Data Mining, Machine Learning, and Data-Driven Predictions for Materials Science and Engineering*, pages 235–270. Springer International Publishing.

[70] Laurene V. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall.

[71] R Marín-Delgado, X Moya, and G G Guzmán-Verri. Landau theory of barocaloric plastic crystals. 6(3):035003.

[72] Justin M. Turney, Andrew C. Simmonett, Robert M. Parrish, Edward G. Hohenstein, Francesco A. Evangelista, Justin T. Fermann, Benjamin J. Mintz, Lori A. Burns, Jeremiah J. Wilke, Micah L. Abrams, Nicholas J. Russ, Matthew L. Leininger, Curtis L. Janssen, Edward T. Seidl, Wesley D. Allen, Henry F. Schaefer, Rollin A. King, Edward F. Valeev, C. David Sherrill, and T. Daniel Crawford. Psi4: an open-source *ab initio* electronic structure program. 2(4):556–565.

[73] Alston J. Misquitta, Anthony J. Stone, and Farhang Fazeli. Distributed multipoles from a robust basis-space implementation of the iterated stockholder atoms procedure. 10(12):5405–5418.

[74] Timothy C. Lillestolen and Richard J. Wheatley. Redefining the atom: atomic charge densities produced by an iterative stockholder approach. (45):5909.

[75] Clare F. Macrae, Ioana Sovago, Simon J. Cottrell, Peter T. A. Galek, Patrick McCabe, Elna Pidcock, Michael Platings, Greg P. Shields, Joanna S. Stevens, Matthew Towler, and Peter A. Wood. Mercury 4.0: from visualization to analysis, design and prediction. 53:226–235.

[76] R. S. Mulliken. Electronic population analysis on LCAO–MO molecular wave functions. i. 23(10):1833–1840.

[77] Alan E. Reed, Robert B. Weinstock, and Frank Weinhold. Natural population analysis. 83(2):735–746.

[78] Shintaro Takenaga, Yoshihiko Ozaki, and Masaki Onishi. Practical initialization of the nelder–mead method for computationally expensive optimization problems. 17(2):283–297.

[79] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

[80] Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. A review on genetic algorithm: past, present, and future. 80(5):8091–8126.

[81] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in python. 17(3):261–272.

[82] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. 32(10):2319–2327.

[83] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. 71(1):361–390.

[84] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. 87(18):184115.

[85] Donald E. Williams. Improved intermolecular force field for molecules containing h, c, n, and o atoms, with application to nucleoside and peptide crystals. 22(11):1154–1166.