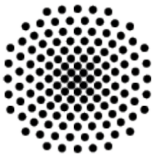
Physics-Driven Machine Learning: from Biomolecules to Crystals



Universität Stuttgart

SimTech

Von der Fakultät Energie-, Verfahrens und Biotechnik der Universität Stuttgart und dem Stuttgart Center for Simulation Science (SC SimTech) zur Erlangung der Würde eines Doktors der Ingenieurwissenschaften (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von

Ángel Díaz Carral

aus Ávila, Spanien

Hauptberichter : Prof. Dr. rer. nat. Dr. h. c. Siegfried Schmauder

Mitberichter : Prof. Dr. rer. nat. Maria Fyta

Mitberichter : apl. Prof. Dr.-Ing. habil. Niels Hansen

Tag der mündlichen Prüfung: 27.03.2024

Institut für Materialprüfung, Werkstoffkunde und Festigkeitslehre der Universität Stuttgart

September 2024

I dedicate this thesis to my loving wife Kany, parents Ángel and Marilo, brother Nacho, and friends. I also extend this dedication to my grandparents, whose wisdom echoes through generations, and to my precious baby Lorán, whose arrival has filled our hearts with boundless joy and love. May the lessons from both the past and the future inspire the discoveries within these pages ...

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. I also declare that there is no conflict of interest, and any included publications are my own work, except where indicated throughout the thesis and summarised and clearly identified on the declarations page of the thesis.

Ángel Díaz Carral

Acknowledgements

This thesis is a compilation of my research work done at SC SimTech and University of Stuttgart during the years 2019-2023.

I am profoundly grateful to Prof. Fyta for her steadfast guidance and the life-changing opportunity to start as a HiWi and progress to a PhD student. Prof. Schmauder's invaluable mentorship played a crucial role during my research journey. Special thanks to Christian Holm for exceptional leadership, unwavering support, and facilitating the contract extension for my continued involvement. The collaborative environment at the Institute for Computational Physics (ICP) and the Institute for Materials Testing, Materials Science, and Strength of Materials (IMWF) was pivotal in developing this thesis. The infrastructure at ICP and IMWF played an instrumental role in its completion.

Expressing gratitude to the dedicated members of my research group at both ICP and IMWF, including Chandra, Takeshi, Martin, Magnus, Simon, Azade, Kira, Ayberk and Louis from ICP, and Xiang, Frank and Stephen from IMWF. Their insights and collaboration significantly shaped the outcomes of this thesis.

I extend thanks to the University of Stuttgart for fostering an enriching academic environment and providing essential resources. My deep gratitude also goes to the University's EXC 2075 SimTech Cluster, supported by the German Research Foundation (DFG), for their financial assistance, allowing me to concentrate on research and pursue academic goals throughout my doctoral studies.

Finally, my deepest gratitude goes to my wife, family, and friends for their unwavering support, encouragement, and understanding throughout this challenging journey. Their love and encouragement have been a constant source of motivation.

This thesis is dedicated to all those who have played a significant role in my academic and personal growth. Thank you for being part of this incredible journey with me.

Ángel Díaz Carral

Publications

The following publications are the result of the work dedicated to this thesis.

- **Á. Díaz Carral**, M. Ostertag, and M. Fyta, "Deep learning for nanopore ionic current blockades," *J. Chem. Phys.*, vol. 154, no. 4, p. 044111, 2021.
- **Á. Díaz Carral**, X. Xu, S. Gravelle, A. YazdanYar, S. Schmauder, and M. Fyta, "Stability of binary precipitates in Cu-Ni-Si-Cr alloys investigated through active learning," *Mater. Chem. Phys.*, vol. 306, p. 128053, 2023.
- **Á. Díaz Carral**, M. Roitegui, and M. Fyta, "Interpretably learning the critical temperature of superconductors: electron concentration and feature dimensionality reduction," *APL Mater.*, vol. 12, p. 041111, 2024.
- **Á. Díaz Carral**, M. Roitegui, A. Koc, M. Ostertag, and M. Fyta, "Concurrent analysis of electronic and ionic nanopore signals: blockade mean and height," *Nano Ex.*, vol. 5, no. 2, p. 025020, 2024.
- **Á. Díaz Carral**, S. Gravelle and M. Fyta, "*In silico* evidence of metastable quaternary phases in Cu-Ni-Si-Cr alloys", submitted APL mach. learn, 2024.
- **Á. Díaz Carral**, M. Roitegui, and M. Fyta, "Structural and electronic features for the prediction of superconducting materials," in preparation, 2024.

Other publications:

- L. Oberer, **Á. Díaz Carral**, and M. Fyta, "Simple Classification of RNA Sequences of Respiratory-Related Coronaviruses," *ACS Omega*, vol. 6, no. 31, pp. 20158-20165, 2021.

Abstract

Physical systems and their interactions are inherently equivariant [1]. The prediction of quantities in machine learning (ML) that are fundamentally generated from these equivariant interactions is accomplished through two main approaches: applying equivariant operations to generate invariant scalar features as input for an invariant model, or utilizing equivariant models themselves. In this thesis, the focus lies on the former framework, where we explore feature extraction and data representation techniques in physics domains through physics-driven machine learning (PDML). This particular field of ML benefits from prior knowledge of physics to create descriptors that encode the underlying symmetries of the dataset, thereby reducing dimensionality, increasing interpretability, and enhancing generalization capabilities. To highlight the significance of physics-driven descriptors in physically-inspired embedding spaces, the thesis focuses on several schemes relevant to its objectives:

1. Copper-based alloys
2. Nanopore detectors
3. High-T_c superconductivity

Through molecular simulations and PDML approaches, the aim is to investigate and provide insights into nanopore sequencing and materials discovery. In this thesis, the following questions are investigated:

- What are the limitations of physics-inspired descriptors in ML?
- Can we decrease the dimensionality of the data while maintaining the same level of prediction accuracy?
- Is it possible to achieve comparable performance using PDML with invariant descriptors compared to conventional ML methods?
- How does PDML scale in atomistic systems?

The investigation of copper-based alloys carried out within the framework of this project focuses on the combination of computer simulations and active learning (AL) to reveal stable precipitate phases of copper alloys and study their mechanical properties. Once the AL cycle that generates accurate ML interatomic potentials (MLIPs) has been successfully implemented, the focus has recently been placed on the stability analysis framework for binary copper-based systems. This part involves quantum mechanical (QM) simulations of various alloy configurations in copper alloys using the density functional theory (DFT) implemented in VASP. Static calculations are performed at zero temperature to generate data for the AL on-the-fly relaxation algorithm. The algorithm utilizes moment tensor potentials (MTPs), a type of descriptors based on invariant polynomials, to construct MLIPs for multi-component alloys. The goal is to conduct a comprehensive search for stable precipitate phases in copper alloys.

To further elucidate the analysis, nanopore DNA translocations are studied through PDML. DNA molecules can be electrophoretically driven through a nanoscale opening in a material, giving rise to rich and measurable ionic current blockades. In this work, ML models are trained on experimental ionic blockade data from DNA nucleotide translocation through 2D pores of different diameters. The aim of the resulting classification is to enhance the read-out efficiency of nucleotide identity, providing pathways toward error-free sequencing. A novel method is proposed that simultaneously reduces the current traces to a few physical descriptors and trains low-complexity models, thus reducing the dimensionality of the data. Each translocation event is described by four features, including the height of the ionic current blockade.

Exploring the field of high critical temperature (high-T_c) superconductivity, an exceptionally effective PDML model is proposed to predict critical temperatures of superconductors by carefully extracting characteristics from electronic and atomic properties. Despite a streamlined feature space, it upholds accuracy when compared to intricate methodologies. The model is fine-tuned to forecast distinct superconductor properties, finding an equilibrium between precision and simplicity and enabling projections for emerging structures.

By bridging the gap between ML and physics, this research will contribute to the growing field of PDML embedding the physics into the descriptors, advancing the ability to model, predict, and control complex physical systems with unprecedented accuracy and efficiency. Through this work, the aim is to pave the way for transformative applications, insights, and discoveries that have the potential to reshape scientific and technological advancements across multiple disciplines.

Zusammenfassung

Physikalische Systeme und ihre Wechselwirkungen sind von Natur aus äquivariant [1]. Die Vorhersage von Größen in maschinellem Lernen (ML), die grundsätzlich aus diesen äquivarianten Wechselwirkungen generiert werden, erfolgt durch zwei Hauptansätze: Anwendung äquivarianter Operationen zur Erzeugung invarianter skalierter Merkmale als Eingabe für ein invariantes Modell oder Verwendung äquivarianter Modelle selbst. In dieser Arbeit liegt der Fokus auf dem ersten Ansatz, in dem wir die Extraktion von Merkmalen und die Darstellung von Daten in physikalischen Domänen durch physikgetriebenes maschinelles Lernen (PDML) untersuchen. Dieses spezielle Gebiet des ML profitiert von dem vorhandenen physikalischen Wissen, um Deskriptoren zu erstellen, die die zugrunde liegenden Symmetrien des Datensatzes kodieren. Dadurch wird die Dimensionalität reduziert, die Interpretierbarkeit erhöht und die Fähigkeit zur Verallgemeinerung verbessert. Um die Bedeutung von physikgesteuerten Deskriptoren in physisch inspirierten Einbettungsräumen zu unterstreichen, konzentriert sich die Arbeit auf mehrere Schemata, die für ihre Zielsetzungen relevant sind:

1. Kupferbasierte Legierungen
2. Nanopore-Detektoren
3. Hoch-T_c-Supraleitung

Durch molekulare Simulationen und PDML-Ansätze soll das Ziel verfolgt werden, Einblicke in die Nanoporensequenzierung und die Materialentdeckung zu untersuchen und zu liefern. In dieser Arbeit werden die folgenden Fragen untersucht:

- Was sind die Einschränkungen von physikbasierten Deskriptoren im maschinellen Lernen?
- Können wir die Dimensionalität der Daten verringern, während wir gleichzeitig die gleiche Vorhersagegenauigkeit beibehalten?

- Ist es möglich, vergleichbare Leistung mit PDML und invarianten Deskriptoren im Vergleich zu konventionellen ML-Methoden zu erreichen?
- Wie skalierbar ist PDML in atomistischen Systemen?

Die Untersuchung von kupferbasierten Legierungen, die im Rahmen dieses Projekts durchgeführt wird, konzentriert sich auf die Kombination von Computersimulationen und aktivem Lernen (AL), um stabile Ausscheidungsphasen von Kupferlegierungen aufzudecken und ihre mechanischen Eigenschaften zu untersuchen. Sobald der AL-Zyklus erfolgreich implementiert und der genaue ML-Interatom-Potentiale (MLIPs) generiert war, wurde der Fokus auf den Stabilitätsanalyserahmen für binäre kupferbasierte Systeme konzentriert. Dieser Teil umfasst quantenmechanische (QM) Simulationen verschiedener Legierungskonfigurationen in Kupferlegierungen unter Verwendung der Dichtefunktionaltheorie (DFT), die in VASP implementiert ist. Statische Berechnungen werden bei Nulltemperatur durchgeführt, um Daten für den AL-On-the-Fly-Relaxationsalgorithmus zu generieren. Der Algorithmus verwendet Moment-Tensor-Potentiale (MTPs), eine Art von Deskriptoren auf der Grundlage von invarianten Polynomen, um MLIPs für Mehrkomponentenlegierungen zu konstruieren. Das Ziel ist es, eine umfassende Suche nach stabilen Ausscheidungsphasen in Kupferlegierungen durchgeführt zu werden.

Um die Analyse weiter zu vertiefen, werden mittels PDML Nanopore DNA-Translokationen untersucht. DNA-Moleküle können elektrophoretisch durch eine nanoskalige Öffnung in einem Material bewegt werden, wodurch reichhaltige und messbare Blockaden des ionischen Stroms entstehen. In dieser Arbeit werden maschinelle Lernmodelle mit experimentellen Daten zur ionischen Blockade von DNA-Nukleotid-Translokationen durch 2D-Poren unterschiedlicher Durchmesser trainiert. Das Ziel dieser Klassifizierung besteht darin, die Effizienz der Nukleotid-Identifikation zu verbessern und so den Weg zu fehlerfreiem Sequenzieren zu ebnen. Es wird eine neuartige Methode vorgeschlagen, die gleichzeitig die Stromspuren auf wenige physikalische Deskriptoren reduziert und Modelle mit geringer Komplexität trainiert, um die Dimensionalität der Daten zu verringern. Jedes Translokationsereignis wird durch vier Merkmale beschrieben, einschließlich der Höhe der ionischen Stromblockade.

Die Erkundung des Feldes der Hochtemperatursupraleitung (high-T_c) schlägt ein außergewöhnlich effektives PDML-Modell vor, um die kritischen Temperaturen von Supraleitern vorherzusagen, indem sorgfältig Merkmale aus elektronischen und atomaren Eigenschaften extrahiert werden. Trotz eines vereinfachten Merkmalsraums behält es seine Genauigkeit im Vergleich zu komplizierten Methoden bei. Das Modell wird

feinabgestimmt, um verschiedene Eigenschaften von Supraleitern vorherzusagen, indem es ein Gleichgewicht zwischen Präzision und Einfachheit findet und Prognosen für aufkommende Strukturen ermöglicht.

Durch die Überbrückung der Kluft zwischen ML und Physik wird diese Forschung zum wachsenden Bereich der PDML beitragen, indem sie die Physik in die Deskriptoren einbettet und die Fähigkeit verbessert, komplexe physikalische Systeme mit einem beispiellosen Maß an Genauigkeit und Effizienz zu modellieren, vorherzusagen und zu kontrollieren. Durch diese Arbeit soll der Weg für transformative Anwendungen, Erkenntnisse und Entdeckungen geebnet werden, die das Potenzial haben, wissenschaftliche und technologische Fortschritte in verschiedenen Disziplinen neu zu gestalten.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

ML techniques and data-driven artificial intelligence (AI) frameworks have seen a remarkable progress in recent decades regarding several application domains with complex, high-dimensional, unstructured data [13]. Despite this success, a significant limitation remains: most ML approaches struggle to extract interpretable information and knowledge from data [14]. Additionally, predictions solely derived from data-driven models may lack physical consistency or even seem implausible [15]. To address these challenges, there is currently extensive research focused on incorporating existing expert knowledge into ML design, so-called knowledge-driven ML (KDML) [16–18]. By carefully integrating into the learning procedure prior domain knowledge about a process, such as physical, biological and/or chemical insights, not only can the quality of the learned representation be improved, but the learning process can also be expedited with fewer data samples [19, 20]. This approach aims to combine the strengths of data-driven techniques with the valuable insights provided by existing knowledge, ultimately enhancing the interpretability and reliability of ML models.

In this context, the main focus lies within the scientific domain of physics as the emerging field of physics-driven machine learning (PDML) is explored. PDML represents the convergence of physics and ML, and its applications are aimed to be explored, along with the seamless integration of physics principles into ML models being investigated.

By incorporating physics-based biases, the goal is to enhance the learning process and improve the performance and interpretability of ML algorithms. These biases can be classified into three types based on where the physics knowledge is embedded: observational, inductive, and learning bias [6]. Observational biases use data that embody the underlying physics or data augmentation techniques to train the machine learning system. Inductive biases incorporate prior assumptions into the model architecture

to ensure compliance with physical laws. Learning biases are introduced through the choice of loss functions, constraints, and inference algorithms during training to favor solutions adhering to physics. These biases can be combined to create hybrid approaches for building physics-driven learning machines. Here, the main focus lies on addressing observational bias by utilizing data transformation and incorporating physical symmetries into invariant physics-inspired descriptors [1]. This approach offers conceptual simplicity in implementation and interpretation compared to equivariant ML architectures or physical loss functions. While it may have limitations when dealing with large volumes of data, invariant descriptors have shown effectiveness in capturing the underlying physical principles of the system, even in high-dimensional feature spaces [6].

In this thesis, the applications of PDML to diverse schemes are divided into two main categories: solid-state physics and biomolecules. In the field of solid-state physics, the focus is on high-throughput screening in materials discovery, specifically high-entropy alloys (HEAs) such as copper-based alloys. The challenges associated with screening a large number of potential candidates are intended to be addressed by employing PDML techniques. Additionally, the superconducting critical temperature of HEAs, including iron-based, cuprate and perovskite superconductors, is investigated using electronic-derived features and low-dimensional surrogate ML models. In the field of biomolecules, novel physical descriptors are introduced for the classification of DNA translocations during experiments with nanopore detectors. This involves developing PDML approaches that utilize the unique characteristics of DNA molecules as they pass through the nanopore, enabling accurate classification and analysis. In the appendix, a highly efficient PDML classification method tailored for respiratory-related coronaviruses is introduced. This innovative approach leverages open reading frames (ORFs) and the genetic code to craft biologically inspired features from the RNA, enhancing virus classification precision. It exemplifies the pragmatic application of PDML in a distinct virology context, significantly enhancing the thesis. Through these diverse applications, this thesis demonstrates the versatility and effectiveness of PDML in various domains, showcasing its potential to revolutionize materials discovery and biomolecular analysis.

1.1 Goals

In this thesis, the aim is to explore the limits of physics-inspired descriptors in ML applied to several atomistic systems. The goals are:

- Improve high-throughput searching in materials discovery: Develop computational methods and techniques to enhance the efficiency and effectiveness of screening processes. This includes analyzing large datasets to identify potential candidates with desired properties, accelerating the discovery and development of materials.
- Enhance read-out protocols in nanopore detectors: Refine and optimize the protocols used for reading DNA molecules in nanopore experiments. Improve accuracy, reliability, and speed to make nanopore detection more precise and applicable in scientific and medical fields.
- Expand the comprehension of high-T_c superconductivity: Integrate PDML techniques with *ab initio* descriptors to predict critical temperatures in novel superconductors. Rely exclusively on features derived from electronic and atomic structures, leading to reduced dimensionality, improved interpretability, and establishing a direct connection between electronic orbitals and heightened prediction accuracy.
- Establish a framework of PDML with invariant descriptors: Create a comprehensive framework that integrates physics principles into ML. Utilize invariant descriptors to capture fundamental symmetries and invariances in physical systems, effectively embedding physics-based knowledge and insights into ML algorithms.

1.2 Thesis Outline

An outline of the thesis is presented as follows:

- In Chapter 2, the state of the art in copper-based alloys, nanopore detectors, high-Tc superconductivity and respiratory-related coronaviruses is explored.
- In Chapter 3, an outline of the ML algorithms applied in physics within this thesis is provided.
- In Chapter 4, a detailed overview of the simulation methods used in this thesis is described.
- In Chapters 5, 6, and 7, the application of PDML to molecules and crystals is presented. Chapter 5 explores the reliability and effectiveness of MLIPs in the search for novel phases within copper-based alloys. In Chapter 6, data acquired from nanopore experiments is utilized to classify DNA molecules and

single nucleotides during nanopore translocations. Various ML models, including clustering and deep learning (DL) methods, are employed for this purpose. In Chapter 7, a highly efficient predictive model designed for estimating the critical temperature of superconductors is introduced.

- In Chapter 8, the contributions of the research and findings are summarized.
- Appendix A presents a supplementary study on fast and efficient approaches for classifying respiratory-related coronaviruses using PDML.

The thesis also includes a References section, where all cited works are listed.

Chapter 2

Theoretical Background

In this chapter, the main concepts and theoretical background of the three chosen fields to which PDML is applied are presented. These fields are related to copper-based alloys, nanopore detectors, high-T_c superconductors, and respiratory-related coronaviruses.

2.1 Copper-based Alloys

Copper-based alloys are of great interest for electric and electronic applications such as connectors and lead frames due to their excellent electrical conductivity and strength [21]. The next generation of integrated circuits requires high-performance copper alloys with high alloy density, multi-functionality, miniaturization, and low cost [22]. These alloys should possess both high electrical conductivity and strength. However, the presence of impurities in the matrix reduces electrical conductivity [23–25]. Thermal aging can lead to precipitation processes, reducing the number of dissolved impurities [26]. Particularly promising alloys include copper (Cu) - nickel (Ni) - silicon (Si) - chromium (Cr) (Cu-Ni-Si-Cr) complexes, which act as effective barriers to dislocation motion, thereby enhancing the alloy's strength [27–29]. Binary systems like Cu-Si, Ni-Si, Cr-Si, Cu-Ni, and Cu-Cr also hinder dislocation movement due to the formation of clusters and intermetallic phases within the Cu matrix [30–34]. Understanding and improving the mechanical and electrical properties of Cu-Ni-Si-Cr alloys during aging rely on the observation and discovery of intermetallic phases. While several stable phases have been experimentally observed, expanding the range of stable structures is important for simulations. Important phases that have been experimentally observed in the Cu-Ni-Si-Cr system and influence its strengthening include Cu_3Si, $Cu_{33}Si_7$, Ni_3Si, Ni_2Si, Ni_3Si_2, Cr-rich clusters, Cr_3Si, Cr_5Si_3 and fcc (Ni, Cr, Si)-rich phase [35–38].

Among these, copper-silicon based alloys are extensively researched for their wide range of applications in high-temperature conditions and microelectronics [39]. They are also used in the synthesis of ultrapure silicon, utilizing the binary precipitate Cu_3Si and its phases μ, μ', and μ" [40, 41]. Cu-rich silicide phases play a significant role in various technical applications, including Li-ion batteries [42]. Studies on the thermodynamics and kinetics of phase formation in the Cu-rich region have utilized dynamic scanning calorimetry and in situ high-resolution transmission electron microscopy. These studies have identified binary precipitates such as Cu_3Si, $Cu_{15}Si_4$, $Cu_{33}Si_7$, Cu_7Si, and Cu_5Si in copper silicides [43–45, 40, 46]. The phase $Cu_{15}Si_4$ and other compounds like $Cu_{19}Si_6$, $Cu_{56}Si_{11}$, and $Cu_{33}Si_7$ are considered stoichiometric due to their narrow ranges of homogeneity [47]. Nickel silicides are commonly used as contacts in electronic devices [48]. First principles calculations on nickel-silicon systems have identified stable phases such as Ni_3Si, $Ni_{31}Si_{12}$, Ni_2Si, Ni_3Si_2, NiSi, $NiSi_2$, and $NiSi_3$, which are also experimentally observed [49–51]. Despite nickel's ferromagnetic nature, no magnetic order has been found in these intermetallic stable phases [52]. Chromium silicides, such as Cr_3Si, Cr_5Si_3, CrSi, and $CrSi_2$, are important transition metal silicides within the studied system. These compounds are of particular interest for their potential use in high-temperature structures [53–57].

Cr-Ni alloys, which are nickel-based alloys, are being considered for potential applications in high-level nuclear waste containers [58–61]. The precipitation of $CrNi_2$ structures requires an exceptionally slow cooling process. Cupronickel alloys, specifically Cu-Ni binary complexes in copper matrices, exhibit promising properties for marine applications and in corrosion environments [62]. They can take the form of nanoparticles [63–65] or clusters [66]. While experimental results have not shown evidence of intermetallic phases in the Cu-Ni system based on the enthalpy of formation [67], cluster expansion methods have identified two prototypes, Cu_7Ni and Cu_8Ni, at low nickel concentrations [68]. The Cu-Cr system exhibits a simple eutectic phase diagram, with Cr-rich clusters present in the microstructure [69, 70]. These clusters are found in the Cu matrix of Cu-Ni-Si-Cr alloys and contribute to their strengthening. No intermetallic phases have been reported in this system based on both first principles calculations and experiments [71–73]. More complex compounds found in the ternary Cu-Ni-Si [74, 75] or Ni-Si-Cr [76, 77] complexes play an important role in the design of copper-based alloys for simulations. Finally, the 4-nary Cu-Ni-Si-Cr system has been studied, where fcc-rich regions in a Cu matrix [38] have been found along with the well known Cr-Si and Ni-Si binary precipitates [78].

The complexity of copper alloys, impurities, and their structures presents ample opportunities for further investigation and the discovery of new materials. *In silico* prediction of novel structures is now possible through advances in computational materials research. Material databases such as AFLOWLIB [79], Materials Project [80], ICSD [81] and OQMD [82] employ density functional theory (DFT) and high-throughput techniques [83] like cluster expansion [84] (CE) and chemical similarity [85] to calculate material properties of intermetallic systems. While QM algorithms show promise, their computational demands hinder their application in high-throughput material searches. CE has been successful in predicting ground-state energies of metallic alloys [86], but its on-lattice nature limits its generalization. Empirical interatomic potentials rely on system-specific parameterization and often lack transferability. The combination of ML approaches, invariant atomistic descriptors, and data from QM simulations has led to the development of MLIPs [87]. They overcome the limitations of classical methods by utilizing complex functional forms that map the potential energy surface (PES) of a system. MLIPs are crucial for accelerating high-throughput searches for novel materials, as they outperform computer simulations in terms of prediction accuracy and computational efficiency. This is particularly important due to the lack of suitable potentials with first-principles accuracy [88–90].

Towards the high-throughput search of materials, different descriptor-based ML methods along with MLIPs have been developed [91, 92]. These models replace *ab initio* calculations by mapping the crystal structure, atomic positions, forces, and stresses to the total energy of the system. Some of the novel descriptors, in combination with regression methods, are the many-body tensor representation [93] with kernel ridge regression, the smooth overlap of atomic positions with gaussian process regression [94] and the moment tensor potentials (MTPs) [95]. The latter are developed based on a polynomial regression. In the field of high-throughput screening, AL approaches [9] in combination with descriptors such as the MTPs have been implemented in order to accelerate the relaxation process of typically very large pool of structure candidates.

In Chapter 5 of this thesis, the focus is on improving the understanding of binary complexes in copper alloys by examining novel stable phases. Additionally, the metastability of quaternary phases is investigated to gain insights into the fcc (Ni, Cr, Si)-rich phase identified by experimentalists. To achieve these goals, a comprehensive framework that integrates QM simulations, AL, and materials data libraries is deployed. The objective is twofold: first, to predict and expand the repertoire of stable phases relevant to Cu-Ni-Si-Cr alloys, and second, to generate MLIPs suitable for large-scale atomistic simulations. By employing this framework, the properties of these novel

prototypes, including mechanical and electronic characteristics, can be further explored. The methodology involves conducting QM simulations for a diverse range of binary and quaternary structures. To identify the most energetically favorable candidates, an on-the-fly relaxation AL scheme based on MTPs is employed. This AL scheme iteratively relaxes potential candidates towards their lowest enthalpy of formation. To assess the stability of the resulting structures, their convex hull is analyzed, and the phonon dispersion is examined. Based on these analyses, potentially stable candidates for copper alloys are proposed. This approach enables the uncovering of promising structures and the advancement of the understanding of copper alloy systems.

2.2 Nanopore Detectors

Nanopores are nanometer-sized holes in materials, which can electrophoretically drive biomolecules, such as DNA, RNA, or proteins through [96–98]. The passage of molecules through a pore results in ionic current and/or electronic signals. These current drops, or ionic current blockades, can be used for the detection of the molecules based on reading-out their identity and sequence or on discriminating among different homopolymers [98, 99]. The duration of each current blockade through a nanopore links to the translocation or dwell time of a biomolecule passing through the nanopore [100]. Typically, the current signals from nanopore experiments need to be post-processed in order to be analyzed and provide information on either the molecule type, length or identity.

Post-prosessing the nanopore data typically involves the use of ML algorithms. These attempt to translate the experimental data into base calls [101] through a proper feature extraction and classification [102–105]. To this end, different ML schemes have been used, ranging from Hidden-Markov-based algorithms [106–110] to neural networks (NNs) [111–114]. Such ML techniques can play a vital role in processing and recovering the information in the nanopore with robust statistics by creating automated models. These have shown the potential to improve the detection accuracy of nanopores and automatize the read-out of the DNA nucleobases towards ultra-fast DNA sequencing. To date, algorithms have been trained in order to process real-time long-read length sequencing data from the MinION nanopore device by Oxford Nanopores [115, 116, 113, 117]. In this way, a nanopore device can efficiently identify, for example, the position and structure of a bacterial antibiotic resistance island [118]. In order to improve error rates in the nanopore data, DL methods such as recurrent neural networks (RNNs) have been implemented [113]. A very important aspect in the

analysis of nanopore data is the classification based on knowledge-based descriptors [119]. Typically, the dwell time or the mean current blockade pointing to the most probable DNA translocation paths are chosen [120, 121]. The most probable DNA translocation paths are typically obtained by considering features such as the dwell time and blockade current [120, 121]. Nevertheless, it has recently been shown that the feature 'dwell time' is quite inefficient in clustering the different types of molecular events through a nanopore [122]. On a time scale lower than the dwell time, important information on molecular aspects cannot be accessed. Instead, the appropriateness and efficiency of another ionic blockade feature, the height, have been demonstrated [122]. Based on an unsupervised clustering of the nanopore data, this feature has been shown to clearly identify specific types of molecular events through the nanopore. Apart from the unsupervised clustering approach utilized in the study [122] and *de novo* clustering [123], most ML algorithms are primarily based on supervised learning. The latter is focused on either improving the algorithmic scaling when processing the data or guiding the learning process to optimize the feature space and reduce the error rates.

Chapter 6 introduces a read-out protocol based on PDML, connecting unsupervised and supervised learning techniques through physics-inspired features. This protocol is designed to address the challenges posed by novel nanopore sequencers, aiming to achieve error-free identification and detection of molecules passing through the nanopore. The feasibility of training ML algorithms using experimental nanopore data within a low-dimensional feature space is explored to enhance interpretability. Furthermore, concurrent ionic and transverse currents from nanopore experiments are jointly processed and analyses in order to understand the importance of these measurements and their inherent details mapped on the choice of features. The primary objective is to identify the key factors that contribute to an improved detection process by minimizing training errors. To accomplish this, an efficient encoding technique for the input data is proposed, emphasizing the significance of data transformation during the pre-processing stage. Implementing these strategies can significantly enhance the data pipeline, accelerate DNA classification, and reduce read-out errors. By employing the PDML read-out protocol, the aim is to enhance the interpretability of ML algorithms trained on experimental nanopore data. This approach holds great potential for advancing the field of nanopore sensing, enabling more accurate and reliable identification and sequencing of molecules translocating through the nanopore.

2.3 High-T_c Superconductivity

The phenomenon of superconductivity in many materials still lacks comprehensive theoretical support. Researchers have historically relied on empirical guidelines, such as Matthias Rules [124, 125], derived from experimental observations due to the absence of theory-based predictive models. These guidelines have aided the synthesis of superconducting materials. However, a major challenge in the field is the discovery of new candidates that exhibit superconductivity at different temperatures, including those with high critical temperatures (referred to as high-T_c superconductors or HTS) [126]. The fundamental mechanisms underlying superconductivity and the relationships between chemical/structural properties and T_c in these materials are not yet well understood [127]. This knowledge gap offers an opportunity for the exploration of novel theories and methods, including computational and data analysis techniques.

Recent advancements in ML have facilitated the investigation of the pairing mechanisms responsible for high-T_c superconductivity and the prediction of critical temperature values [128]. Moreover, the availability of extensive materials databases, encompassing both experimental and calculated properties, has enabled the development of advanced data-driven ML approaches to discover potential high-T_c candidates [129]. However, due to the complex nature of superconductivity, interpreting these models remains challenging [130]. Descriptor-based ML approaches, leveraging chemical and structural features, have emerged as a promising avenue [131]. These methods not only predict the critical temperature T_c and identify potential novel superconducting structures but also enable the study of the importance of different descriptors [131]. This insight into the physical concepts underlying superconductivity can inform the development of novel theories. Unsupervised learning techniques, including conventional and neural network-based clustering methods like k-means [132], DBSCAN [133], and Self-Organizing Map (SOM) [134], have been employed to group superconductors into meaningful classes and discover new materials categories [135]. Additionally, Convolutional Neural Networks (CNNs) have been used for feature learning to effectively distinguish between cuprate and iron-based superconductors [129].

Material features, such as elemental property statistics generated through the Materials Agnostic Platform for Informatics and Exploration (Magpie) [136, 137], have shown promise when combined with tree-based ML algorithms [136]. Learned predictors using these features offer potential insights into the mechanisms governing superconducting effects. Other approaches based on features derived from existing databases [138, 128, 139] and/or chemical formulas [140, 141] have revealed that certain predictors are not directly influential in superconductivity and do not significantly

improve the models [130]. The integration of Magpie features and structural information through Smooth Overlap of Atomic Positions (SOAP) descriptors [142, 143] has shown improvements over other methods [144]. Despite notable progress in applying ML techniques to superconductivity, many models still lack sufficient prediction accuracy or rely on high-dimensional feature spaces, which increases learning complexity.

Chapter 7 of the thesis introduces a PDML model that leverages a minimal set of electron-specific features to accurately predict the critical temperature of superconductors. The goal of this model is to achieve high fidelity in predicting superconducting behavior. The first aim is to assess the feature importance of electron-specific descriptors by leveraging unsupervised learning techniques, particularly dimensionality reduction through projection. The objective is to gain insights into the data's clustering potential within an embedding space. Additionally, SHAP (SHapley Additive exPlanations) analysis is employed to further elucidate the contribution of individual features to the clustering outcomes. To validate the effectiveness of the novel method, experiments are conducted using a list of potentially new HEA superconductors. The approach is compared against other relevant methods described in the literature, providing valuable insights into the performance and superiority of the PDML model in predicting the critical temperature of superconducting materials. The efficacy of the approach is demonstrated and compared to existing methods, contributing to the advancement of the field and showcasing the potential of PDML in predicting superconductivity in diverse materials.

2.4 Respiratory-Related Coronaviruses

The coronavirus SARS-CoV-2 has been spreading globally, and efforts are being made to isolate and control its spread [145–148]. Over 22,000 genome sequences have been collected since its identification [149]. Research studies are focused on developing drugs or vaccines using these sequences [150, 151]. Accurate identification and categorization of the virus are crucial for reducing the disease's spread [152]. Algorithmic approaches, such as the UMAP algorithm, have shown promising results in identifying SARS-CoV-2 viruses in genome datasets [153–155]. UMAP is widely used in bioinformatics and clustering visualization. Existing methods for virus identification, such as the one referenced in [156], are computationally complex and unsuitable for smaller computer architectures like microcontroller chips. To enable widespread and easier virus identification, as well as facilitate fast initial identification while reducing complexity, straightforward and efficient approaches are required. One potential

solution is to employ a theory-based approach that leverages the biological information embedded within viruses. Viral proteins are encoded by virus sequences using codons, which are translated into amino acids. These amino acids form proteins. The protein-coding segment is called an ORF [157, 158]. They can uncover overlapping and hidden genes in viruses, including SARS-CoV-2 [159]. An ORF starts with a start codon, contains the protein sequence, and ends with a stop codon. Variations in ORF regions differentiate virus types within a family. The number of substrings, or k-mers, of length k in a sequence is similar among viruses within the same family [160]. Techniques focused on detecting RNA-genome substrings often employ larger k-mer sizes. In addition, some techniques utilize natural vectors to create a vast and detailed space, where each biological molecule is uniquely represented [161]. SARS-CoV-2 and other SARS-type viruses have a large ORF called ORF1ab, spanning about 13,000 nucleobases [162, 163]. ORF1ab contains essential structural proteins for virus replication [164].

In Appendix A, the application of PDML to the prediction of COVID-19 from its RNA is discussed. An efficient approach using genetic code rules and ORFs to encode the entire SARS virus sequence into biological features is proposed. The method offers greater interpretability of variations in RNA codon frequencies, also known as codon bias [165]. To achieve this, the genetic code rules (3-mers) are utilized to construct biology-based features, which is a natural choice. MERS-CoV, SARS-CoV, SARS-CoV-2, and other related viruses are analyzed to identify distinct clusters. By collecting coronavirus family data, extracting features from ORFs and codon counts, and visualizing low-dimensional latent spaces, the goal is to achieve accurate clustering. In order to demonstrate the effectiveness and validate the proposed approach, the complexity and diversity of the analyzed viral RNA data are enriched. Initially, the focus is on SARS-CoV-2, SARS-CoV, and MERS-CoV. Subsequently, the analysis is expanded to include additional members of the coronavirus family. Finally, members from other virus families, such as the herpes DNA virus family, are incorporated. This progressive inclusion of diverse viral data allows for the strengthening and evaluation of the robustness of the approach.

Chapter 3

Machine Learning

In this chapter, an introductory overview is provided for both shallow and PDML methodologies. The focus is on elucidating the key techniques pivotal to this thesis, which include clustering analysis, tree-based methods, DL, feature importance, PDML and MLIPs.

3.1 Shallow Learning

ML involves computers using algorithms to optimize specific performance measures, like character recognition, based on example data or past experiences. It has evolved as a distinct field within computer science since the 1980s, with applications in engineering, speech and image analysis, pattern recognition, and communications [166]. Learning algorithms enhance the efficiency of target algorithms, serving as alternatives to conventional data extraction from simulations. ML's ultimate goal is to enable computers to solve problems without explicit programming, achieved through learning rules that save computational time and improve accuracy beyond human capability. Technologies such as image and voice recognition, personalized marketing, and data analytics work in conjunction with ML algorithms to acquire knowledge and glean valuable insights [167]. Shallow learning algorithms follow a pipeline: acquire, preprocess, and transform data, select relevant features to create a feature space, form a training set, train the algorithm to identify distinct zones, optimize by finding similarities, and create a decision rule. ML algorithms can be categorized into five main classes based on whether prior knowledge is required or the goal is to discover new patterns [168]:

- Supervised learning (SL): In SL [169], samples are assigned classes (categorical or numerical) based on known labels. These algorithms use labeled data to automatically classify new samples. They learn from input datasets to generate outputs, with the classification task aiming to predict labels for new inputs [170]. This is especially valuable in computational biology for predicting mechanisms with uncertain definitions. The general mapping for supervised learning is:

$$Y = f(X, \theta) \tag{3.1}$$

where:

Y is the output or target variable,
X is the input features,
f is the model function capturing the relationship between inputs and outputs,
θ represents the parameters (weights and biases) of the model,

The objective is to find the values of θ that minimize this cost function. The least squares error (LSE) cost function is the most common in supervised learning. The optimization process outlined in Eq. 3.1 requires finding the solution to the equation:

$$\theta = \underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{3.2}$$

where:

m is the number of training examples,
h_θ is the predicted output, often computed using an activation function,
$y^{(i)}$ is the actual output.

For a general activation function denoted as σ, the predicted output $h_\theta(X)$ is calculated as $h_\theta(X) = \sigma(z)$, where z is the linear combination of the input features $X = \{x_1, x_2, ..., x_n\}$ and their corresponding weights θ, in addition to a bias term θ_0:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{3.3}$$

The choice of the activation function (σ) depends on the specific requirements of the model. For example, in the case of logistic regression, the sigmoid function is commonly used as the activation function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.4}$$

The interpretation of $\sigma(z)$ depends on the application, often representing the probability of an input x belonging to a class in classification tasks. The decision boundary is determined by comparing this probability to a threshold. Different models use varied activation functions, impacting the expressiveness and characteristics of the model.

- Semi-Supervised learning (SSL): The goal of these algorithms is to predict unknown labels from a dataset created for classification purposes. A trained supervised algorithm is utilized to classify unlabeled data [171]. The most confident unlabeled samples and their predicted labels are incorporated into the training set.

- Unsupervised Learning (UL): Unlike SL algorithms, UL is employed when the labels are unknown. The training set consists of unlabelled samples, which means there are no predefined classes for dividing the feature space [167]. The purpose of UL is to observe the underlying mechanics of the system and uncover insights by identifying groups of samples with similar features. Some examples of UL algorithms include clustering methods, anomaly detection algorithms, or unsupervised versions of NNs [172]. As example of UL such as the k-means clustering algorithm, given a dataset $\mathbf{X}$ represented by the feature vector $\{x_1, x_2, ..., x_n\}$ in $\mathbb{R}^n$ and k clusters, the goal is to find cluster centers $\mu_1, \mu_2, ..., \mu_k$. The optimization objective is to minimize the sum of squared distances:

$$\underset{\mathbf{C},\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{X_i} - \mu_{c_i}\|^2 \tag{3.5}$$

  Here, $\mathbf{C} = \{c_1, c_2, ..., c_n\}$ represents cluster assignments, and $\mu = \{\mu_1, \mu_2, ..., \mu_k\}$ are cluster centroids. In Fig 3.1, a comparison between the two main ML methodologies is depicted.

- Reinforcement Learning (RL) involves an agent learning an optimal policy through trial and error in interaction with its environment. It's used in various fields

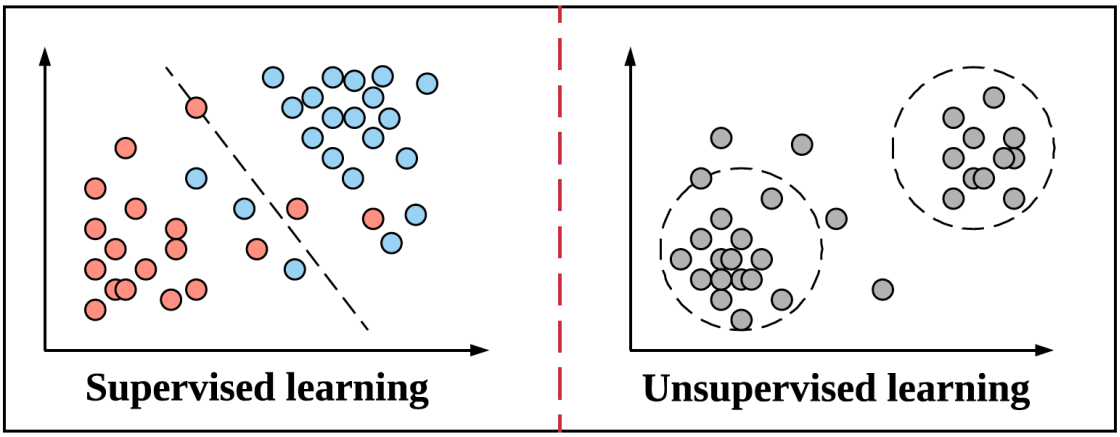


Fig. 3.1 Instances of SL classification (left) and UL clustering (right). In SL, the model is trained on labeled data (blue and red), allowing it to learn patterns and relationships. On the other hand, UL involves the model adapting to unlabeled data (grey), autonomously identifying structures and patterns without predefined categorization [2].

and aims to maximize future rewards by selecting actions based on current environmental states [167]. RL sits between supervised and unsupervised learning, focusing on actions that contribute to a cumulative increase in reinforcement signal values for long-term performance [173].

- Active learning (AL): This specialized area within ML is employed when obtaining labels for a supervised task, such as regression or classification, is costly or resource-intensive. It aims to optimize the training set by actively selecting the most informative training samples that contribute to highly accurate predictions, thereby minimizing the loss function of our model [174].

3.1.1 Clustering Analysis

Clustering analysis categorizes dataset samples into subsets based on similarities [175]. Each subset comprises similar yet distinct samples, positioned according to distances to cluster centroids. Centroids represent central points, with larger distances indicating lower similarity. Clustering methods can be split into four main categories [176]:

- Hierarchical methods: These approaches form a tree-like structure by recursively dividing the dataset. Two types exist: agglomerative, which starts with individual samples and merges clusters iteratively, and divisive, which does the opposite [177].

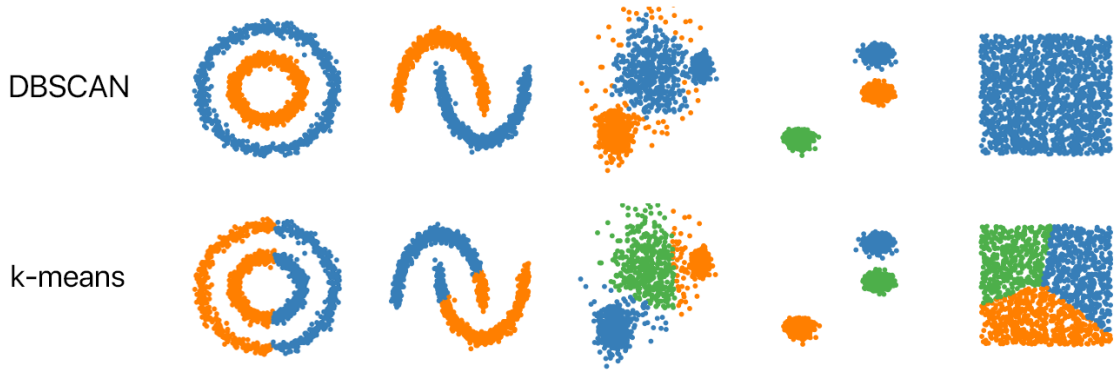


Fig. 3.2 k-means vs. DBSCAN example. DBSCAN proves adept with irregular and diverse datasets, while k-means efficiently partitions data into k clusters based on mean distances to centroids [3].

- Density-based methods: Similar to distance-based clustering methods, these techniques assign samples to clusters based on the concept of density rather than distance. One of the most known density-based methods is DBSCAN [133]. By utilizing density, they can identify clusters with arbitrary shapes, avoiding the assumption of spherical clusters in the feature space. Density-based methods are also effective in detecting outliers or anomalies within the dataset [178].

- Grid-based methods: These algorithms are an improvement over density-based methods and are particularly suitable for datasets with higher dimensions. They quantize the feature space into a finite number of cells using a grid-like data structure. By operating within this grid structure, the clustering algorithm identifies clusters in an efficient manner [179].

- Partitioning methods: commonly used in clustering, divide a dataset into a specified number, k, of clusters using a distance metric. This results in spherical clusters, as seen in k-means clustering, a popular example. The k-means algorithm, an NP-hard method, partitions the dataset into k clusters, with the mean observation in each cluster as the centroid [180]. It requires inputting the desired k value and iteratively produces a cluster representation as output.

Fig 3.2 depicts a comparison between the two clustering methods used in this thesis. After clustering analysis, it's crucial to evaluate the results through clustering validation. This is necessary as clustering algorithms yield results even when the dataset doesn't naturally form distinct clusters. Evaluation becomes essential to measure the

method's effectiveness. Cluster validation can be internal, assessing clustering solution stability, or external, comparing results with other datasets and methods [181]. Key steps in evaluating clustering performance include:

- Determining the clustering tendency of the dataset to identify non-random structure.
- Identifying the appropriate number of clusters.
- Assessing how well the cluster analysis results fit the data independently.
- Comparing results with externally known information, like class labels.
- Comparing two sets of clusters to determine better quality or agreement with known information [182].

3.1.2 Tree-based Methods

Tree-based methods, detailed by [183], are prominent nonparametric models that employ decision trees to iteratively divide a training dataset into smaller, more homogeneous subsets, effectively handling both classification and regression tasks [184]. Each node in the tree is associated with a decision rule, guiding the distribution of the data inherited from its parent among its children, and every leaf node, also referred to as a sub-group, is linked to at least one data point from the original training set. The most common criteria for node splitting are:

- Gini Impurity: Minimize Gini impurity to achieve maximum homogeneity in subsets.
$$I_G(X) = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.6}$$
- Entropy: Maximize information gain to decrease entropy and enhance homogeneity.
$$I_E(X) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{3.7}$$

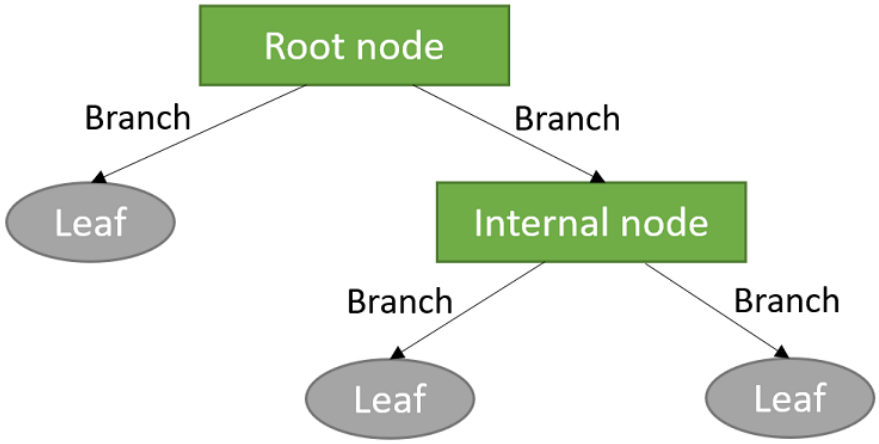


Fig. 3.3 This example employs two features(root and internal nodes) to classify data into three sub-groups (leaves) and intermediate splits, not immediately forming a leaf, are internal nodes. The lines connecting nodes and leaves are branches. Input variables (features) are utilized to classify data into sub-groups based on binary conditions. The training sample computes the sample mean of the output y_t for each sub-group, serving as a constant prediction for future observations classified into that sub-group [4].

where:

X is the dataset or a specific node,

c is the number of classes,

p_i is the proportion of instances of class i in the dataset or node X

In Fig. 3.3, a representation example of a decision tree is depicted. The construction of the tree entails the iterative segmentation of variables, where branches undergo evaluation for accuracy, efficiency, and effectiveness. The subset of variables chosen to split an internal node relies on predetermined criteria formulated as an optimization problem. To enhance the efficiency of the tree, the strategy involves reordering variable splits, giving priority to essential variables at the top, and eliminating irrelevant features for a successful model. Tree-based methods stand out as some of the most robust ML algorithms, adept at accommodating complex datasets and ranking among the most powerful algorithms in use today [185]. They demand minimal preparation time, dispensing with the need for feature scaling or centering. These methods not only yield excellent predictions but also allow you to scrutinize the calculations behind these predictions. However, articulating the reasons behind predictions in simple terms can

be challenging. Despite their tendency to overfit, they consistently outperform DL on tabular data, as highlighted by [186].

Random Forest Algorithms

Random forest (RF) methods evolved from empirical successes rather than from a sound theory, with various parts of the algorithm remain heuristic rather than theoretically motivated [187]. These models combine tree predictors, with each tree drawing values from a random vector independently sampled with the same distribution across all trees [188]. Each decision tree is trained independently of the others and on distinct subsets of the training data. The ultimate decision is reached by considering the more frequently predicted outcome. As the number of trees in the forest grows, the generalization error converges to a limit. The generalization error of a forest of tree classifiers hinges on the strength of individual trees and the degree of correlation among them.

Extreme Gradient Tree Boosting: XGBoost

XGBoost (XBG), an ML technique detailed in [189], utilizes an optimized ensemble model of classification and regression trees. It employs gradient boosting to create a decision tree ensemble for making predictions. Gradient boosting, an ensemble learning method for ML classification and regression problems, combines multiple decision trees to construct a robust model for accurate predictions [190]. The algorithm builds trees sequentially, with each tree aiming to rectify the errors of its predecessor. Noteworthy for its outstanding performance on various standard classification benchmarks, XGB distinguishes itself by running significantly faster than many other popular approaches, as emphasized in [191].

3.2 Deep Learning

The DL paradigm is a sub-field of ML inspired by the structural and functional characteristics of neurons in the human brain [167, 192, 193]. Artificial Neural Networks (ANNs) are utilized to address non-linear regression and classification problems where linear activation functions are insufficient. As problems become more complex, the number of features exponentially increases, necessitating the consideration of linear combinations among them. Traditional linear ML algorithms struggle with the computational cost of such problems. However, the development of more complex architectures

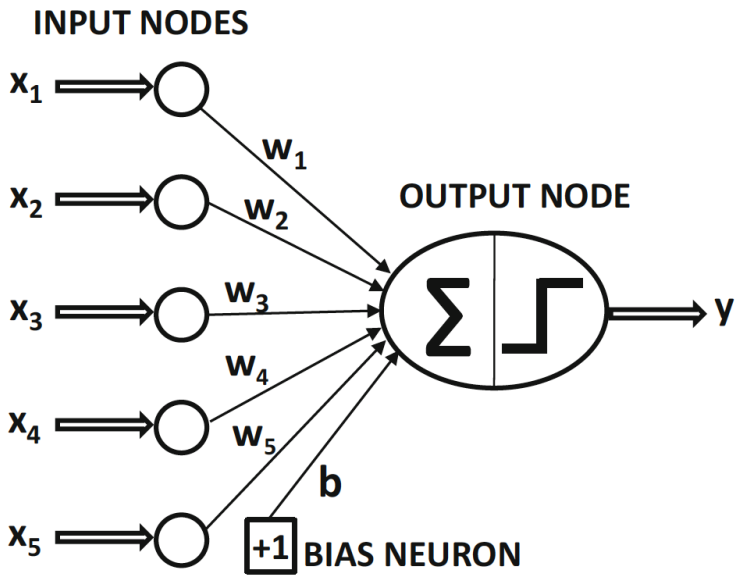


Fig. 3.4 Computational graph of a single perceptron with the input and output layers, as well as the nodes and bias vectors of the layers [5].

has provided a solution. The Single Layer Perceptron, introduced by Rosenblatt, was the first non-linear model [194]. This architecture consist of an unidirectional network formed by one input layer and one output layer. Selecting as activation function the sign function (or the sigmoid function) this algorithm is the most simple NN created.

3.2.1 Multi-Layer Perceptron

Feedforward NNs, commonly known as Multilayer Perceptrons (MLPs), are composed of multiple layers of single perceptrons. These networks consist of an input layer, one or more hidden layers, and an output layer. Each layer contains multiple computational units. Fig. 3.4 illustrates an example of the most simple deep neural network (DNN) or single perceptron, with only one hidden layer. The next terms can be identified:

- **Input Nodes ($\mathbf{x_i}$):** Representing input features, each of the n features corresponds to an input node.
- **Weights ($\mathbf{w_i}$):** Parameters learned during training, W_{ij} denotes the weight connecting input node i to hidden node j.
- **Bias Neuron (b):** An additional parameter for each hidden node, enabling the network to shift activation output.

- **Activation Function (σ):** Applied to the weighted sum of input nodes plus the bias term, common functions like sigmoid, tanh, and ReLU result in node activation a_j:

$$a_j = \sigma\left(\sum_{i=1}^{n} W_{ij} \cdot X_i + b_j\right) \tag{3.8}$$

- **Output Node (y):** Producing the final result, the output node(s) apply an activation function to the weighted sum from the last (hidden) layer.

With an adequate number of units in a single hidden layer and the appropriate activation function, the network can approximate continuous functions arbitrarily closely. However, empirical evidence suggests that deep networks with multiple hidden layers exhibit improved performance and lower generalization error compared to shallow networks with a single hidden layer. With increased computational power, training larger networks in less time becomes feasible. This allows for efficient testing of different structures and hyperparameters. Additionally, larger datasets contribute to better generalization of the network's learned patterns.

3.3 Feature Importance: SHAP Values

SHAP (SHapley Additive exPlanations) is a potent method for establishing a hierarchy related to feature importance. Initially introduced by Scott M. Lundberg and Su-In Lee [195], it has become a widely employed tool among data scientists to explain individual predictions and provide interpretability to model descriptors. The core concept in SHAP analysis involves approximating a given model in an additive way to establish a hierarchy of feature importance. To achieve this, a model is approximated by a function of the type:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.9}$$

where $z' \in \{0,1\}^M$ represents a simplified input, and M is the number of simplified input features. The SHAP values, denoted as $\phi_i \in \mathbb{R}$, are utilized to assess the model output. A mapping function $h_x(x') = x$ relates the simplified inputs approximating the model to the inputs of the actual model. The relationship between the simplified inputs for approximating the model and the inputs of the actual model is given by the mapping function $h_x(x') = x$. This mapping includes information on the actual input, enabling easy interpretability of the approximating model. Thus, the approximating

model results in a linear combination of ϕ values, which are determined through the following form:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!}[f_x(z') - f_x(z'/i)], \tag{3.10}$$

where f refers to the actual model, and $f_x(z') = f(h_x(z'))$ denotes the output of the actual model for the simplified input z'. For each data point x', new artificial data points z' are created by excluding features, and the model is evaluated. The differences in the model output between the original data points and the artificial data points yield the SHAP value for each feature i.

In the ML context, SHAP values account for varying magnitudes and signs, indicating how features contribute to the model's output or class prediction. A positive sign indicates a contribution to predicting a specific class, while a negative sign signifies a contribution to predicting the opposite class. The key advantages of SHAP values lie in their transparency, making them easily understandable. Additionally, this procedure is model-agnostic and can be applied to approximate any model.

3.4 Physics-Driven Machine Learning

Physics-driven ML (PDML), a subset of scientific ML (SciML), represents the synergy between physics and machine learning. The core principles of SciML involve addressing the challenges presented by scientific domain knowledge and developing interpretable and robust ML models and algorithms [196]. The integration of physics knowledge into ML models is expected to improve accuracy, physical interpretability, model size, complexity, sample efficiency and generability [7]. Physics domain knowledge is available in various forms, including essential physical principles (e.g., ab initio or first-principles physics), physical constraints (e.g., symmetries, invariances, conservation laws, asymptotic limits), and valuable insights gained from theoretical or computational studies [14]. In this thesis, the mission is to explore dimensionality reduction of collected observational data through PDML, transitioning from data-driven to physics-driven approaches. To better understand this transition, the interplay between data and physics scenarios in the ML field is illustrated in Fig. 3.5.

The incorporation of physically relevant prior knowledge into ML algorithms can be achieved through various high-level approaches: physics-inspired descriptors, ML architecture, loss function, and the utilization of hybrid methodologies. To incorporate this valuable knowledge into our models, three primary approaches are employed

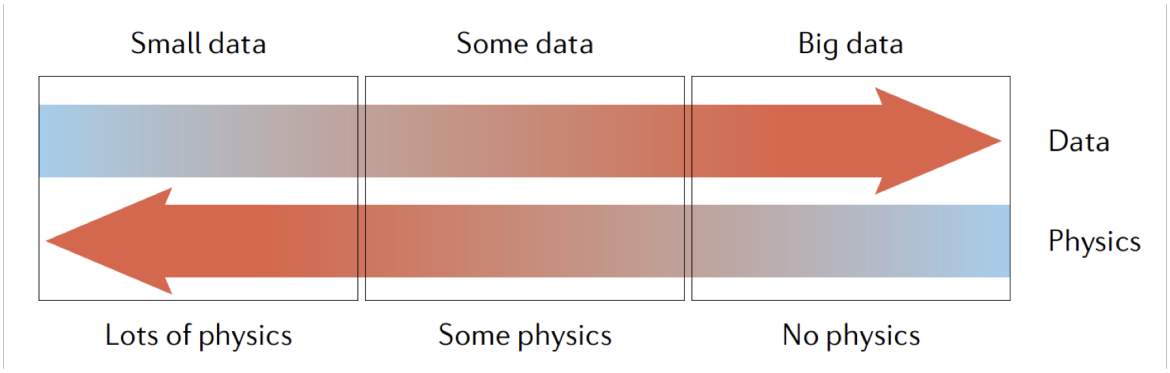


Fig. 3.5 Data and physics scenarios [6].

according to the categorization of bias in the ML process [1, 6]: The first approach ingeniously incorporates observational bias by employing equivariant operations to generate invariant scalar features. These operations carefully preserve the consistency and integrity of the resulting features even under various data transformations, thereby safeguarding crucial physical properties throughout the learning process. By training a ML algorithm on an invariant, scalar, lower-dimensional feature space, it gains the capability to learn functions, vector fields, and operators that faithfully reflect the underlying physical structure of the data. The second approach relies on the concept of inductive bias, accomplished through the utilization of equivariant models. These models offer the distinct advantage of more faithfully representing physical interactions, ensuring that essential quantities maintain their predictable behavior under various coordinate transformations. The third approach tightly constrains the learning optimization step by skillfully incorporating learning bias through the selection of appropriate loss functions, constraints, and inference algorithms during the training phase. These strategic choices are deliberately tailored to steer convergence towards solutions that harmonize with the underlying physical principles. Through the fine-tuning of soft penalty constraints, the model can approximate adherence to the governing physical laws, providing a versatile avenue to introduce a diverse array of physics-based biases. As an illustrative example, we mention Physics-Informed Neural Networks (PINNs). While not directly applied in this thesis, they are pertinent to this section. PINNs represent a SciML technique used to tackle problems related to Partial Differential Equations (PDEs). These networks approximate PDE solutions by reframing the task of directly solving governing equations into an optimization problem centered around a loss function [197]. All these methods are designed to maintain the symmetries and equivariances present in the underlying physical systems, making them well-suited for effectively capturing and representing the relevant information.

By leveraging these three approaches, ML methods can effectively embed physical domain knowledge into their frameworks, enhancing their capabilities and enabling the application of ML techniques in a wide range of physical and scientific disciplines.

3.4.1 Symmetry and Equivariance in ML

Symmetry, in the context of an object or system, refers to a transformation that preserves a specific property, rendering it unchanged or invariant [198]. These transformations can manifest as either smooth, continuous processes or discrete operations. Symmetries play a fundamental role in various ML tasks. Discrete symmetries naturally emerge in scenarios like particle systems, where particles lack a definitive order and can be rearranged arbitrarily. Similarly, they arise in various dynamical systems through concepts such as time-reversal symmetry, as seen in systems adhering to detailed balance principles or Newton's second law of motion. Furthermore, permutation symmetries are of central importance in the analysis of data organized in graph structures. Mathematically, symmetries are typically described using groups [7]. The relationship between a function f and a symmetry group G can be characterized by examining its equivariance properties, indicating that f is equivariant with respect to G. Invariance is a special form of equivariance, dealing with quantities that remain unchanged irrespective of the choice of the coordinate system. Fig. 3.6 illustrates the distinction between invariance and equivariance. In the field of ML, a way to categorize models into PDML or conventional ML is based on whether symmetry is employed or equivariant operations are used [1].

By utilizing an equivariant model, transforming the input results in an output representation that undergoes the same transformation [199]. This often includes incorporating geometric coordinates and relevant quantities crucial for describing the system's behavior, such as external fields or atom-wise properties like velocities. The strength of equivariant models lies in their capacity to uphold the system's symmetries and invariances throughout the learning process, ensuring a robust and accurate representation of the underlying physics. The main concern associated with equivariant ML models revolves around the substantial technical complexity they involve. On the other hand, an invariant function produces the same output for both transformed and non-transformed inputs. Invariant scalar features are preferred over geometric tensors due to their ease of handling and computational efficiency [1]. The application of invariant models has demonstrated impressive performance across numerous existing benchmarks, making them a compelling choice for various scientific and engineering applications. By effectively capturing the essential invariances present in the data,

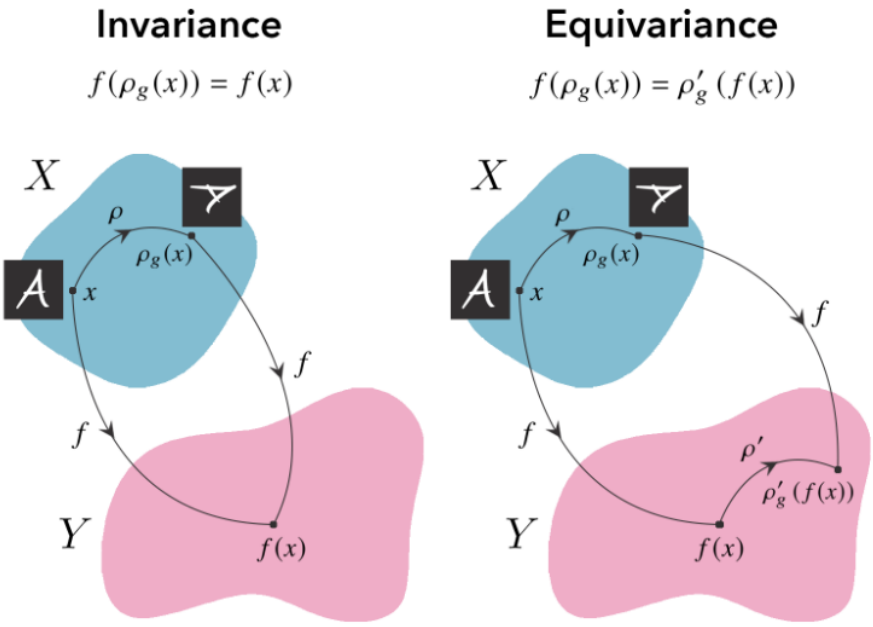


Fig. 3.6 An example illustrating the differences between symmetry group invariance and equivariance is presented in the context of identifying a handwritten letter in an image [7].

these models enable accurate and robust predictions, advancing our understanding and problem-solving capabilities in the domain of physical systems. However, when utilizing an invariant model, the challenge lies in devising a method to represent your naturally equivariant physical system using invariant scalar features. This requires careful consideration and creativity to encapsulate the crucial characteristics of the system in a way that remains consistent under transformations.

It is crucial to note that, even when the ultimate goal is predicting a scalar quantity, not all physical interactions can be adequately represented using scalars alone. The richness and complexity of physical phenomena often necessitate a more comprehensive representation that considers higher-order interactions and geometrical aspects. This is precisely where equivariant models excel. By utilizing invariant and equivariant models, researchers and practitioners can access a powerful toolset to effectively capture the intricate dynamics of physical systems and make accurate predictions across a wide range of scientific and engineering applications.

3.4.2 Invariant Descriptors: Learning Latent Representations

Observational data plays a fundamental role and serves as a critical foundation for the success and recent achievements of ML algorithms [6]. Nevertheless, it is essential to recognize that these data can also inadvertently introduce biases into the learning process. Despite this, ML methods have proven to be remarkably powerful, particularly when provided with sufficient data that cover the entire input domain of a learning task. This capability enables accurate interpolation even in high-dimensional scenarios. It is crucial to ensure that these observational data capture the underlying physical principles governing their generation. By doing so, we can leverage these data as a means of weakly embedding these principles into an ML model during its training phase. Nonetheless, it is worth noting that for over-parameterized ML models, a substantial volume of data is typically required to reinforce these biases adequately. This reinforcement is crucial to generate predictions that respect essential symmetries and conservation laws within the physical systems. Unfortunately, obtaining such a large volume of data can be challenging and costly, especially in the context of physical and engineering sciences. In many cases, observational data may be generated through expensive experiments or large-scale computational models, making the cost of data acquisition potentially prohibitive. As such, researchers and practitioners must carefully consider the trade-offs between the data volume required and the resources available for data acquisition in these applications, as mentioned in Fig. 3.5.

In handling high-dimensional unstructured data, a first step is reducing dimensionality by extracting informative features [200]. These features form the foundation for solving downstream tasks like prediction or classification, boosting efficiency and accuracy in ML tasks. A highly efficient approach to tackle the data acquisition challenge is to create a lower-dimensional tuple of physics-inspired descriptors, generating an embedding or latent spaces that capture all the underlying symmetries of the system. This resultant latent space serves as a robust basis for training an ML model, eliminating the necessity for data augmentation and greatly improving sample efficiency. By leveraging the intrinsic knowledge embedded within these physics-inspired representation, the ML model gains a deeper and more meaningful understanding of the data. This leads to more accurate predictions and maximizes the use of available samples, culminating in highly effective and precise machine learning applications. This integration of physics-driven features not only streamlines the learning process but also enhances interpretability and opens doors for innovative advancements in diverse scientific and engineering domains.

3.4.3 Geometric Deep Learning

Introducing an inductive bias involves the development of specialized architectures that intrinsically incorporate prior knowledge and inherent biases relevant to a specific predictive task [6]. This concept aligns with the paradigm known as geometric deep learning (GDL), which spans the entire spectrum of deep learning, encompassing both Euclidean and non-Euclidean domains. GDL seamlessly integrates insights about the structure and symmetry intrinsic to the system of interest. These domains encompass intricate structures, such as graphs, manifolds, meshes, and string representations [201]. Fundamentally, GDL employs techniques that establish a geometric bedrock, entailing the assimilation of knowledge concerning the inherent spatial relationships and symmetrical attributes present in input variables. By infusing this geometric foundation, the objective is to enhance the precision of information captured by the model [202]. Among these methods, CNNs stand out as a canonical example, fundamentally reshaping the landscape of computer vision by adeptly preserving invariances related to symmetrical groupings and the distributed patterns found in natural images. Furthermore, convolutional networks can be extended to accommodate additional symmetry groups, encompassing rotations, reflections, and more intricate gauge symmetry transformations. Other notable instances include graph neural networks (GNNs), equivariant networks, kernel methods such as Gaussian processes, RNNs and Transformers [198]. GDL provides a constructive procedure for incorporating prior physical knowledge into neural architectures.

Convolutional Neural Networks

CNNs draw inspiration from cognitive neuroscience, particularly the pioneering work of Hubel and Wiesel on the cat's visual cortex. Their research uncovered distinct neuron types: simple neurons responsive to small visual patterns and complex neurons tuned to larger motifs [203]. CNNs serve as the cornerstone of Image Classification, dominating the landscape of Computer Vision algorithms. Moreover, they have found promising applications in Natural Language Processing. In CNNs, the core operation is convolution:

$$f_i = \sigma\left(\sum_j (X \circledast W_{ij}) + b_i\right) \tag{3.11}$$

where filters systematically extract descriptive features by traversing the input data, generating feature maps. These filters act as functions applied to the data. CNNs accommodate multi-dimensional input arrays, such as two-dimensional images with

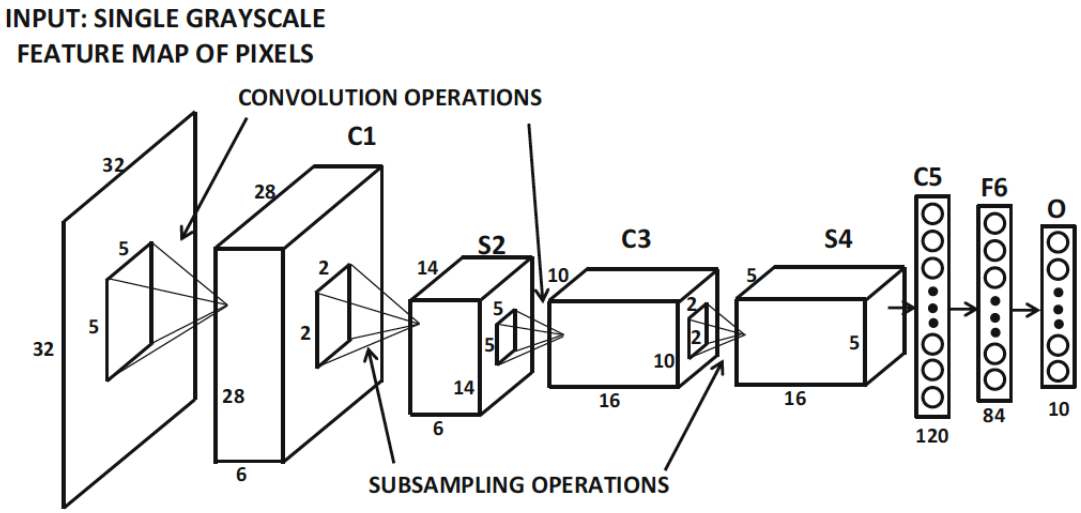


Fig. 3.7 LeNet-5: One of the earliest convolutional neural networks [5].

three color channels or one-dimensional genomic sequences with a channel for each nucleotide [204]. The high dimensionality of images increases the complexity of hyperparameter tuning. Convolutional layers, often referred to as pooling layers, empower the network to autonomously acquire abstract features. A typical CNN architecture is illustrated in Fig. 3.7. Key hyperparameters, including the number of convolutional layers, filter count, and filter size, need fine-tuning during the validation process. ~~s~~CNNs excel at detecting local patterns by employing the convolution operation to glean insights from data. This process involves scanning through the data and generating feature maps that connect to subsequent CNN layers. The input to CNNs is an n-dimensional tensor, representing a variety of data types. For instance, it can encompass two-dimensional images with three color channels or one-dimensional genomic sequences with a channel assigned to each nucleotide. The integration of convolutional and pooling layers enables CNNs to autonomously discover abstract features within the data.

Recurrent Neural Networks

Another type of GDL tailored for processing specific data types, such as time-series, text, and biological data containing sequential dependencies among attributes, is RNNs. Renowned for their proficiency in learning from string representations, RNNs excel in pattern recognition across various time steps, facilitated by parameter sharing across different model segments. In an RNN, there is a direct correspondence between the

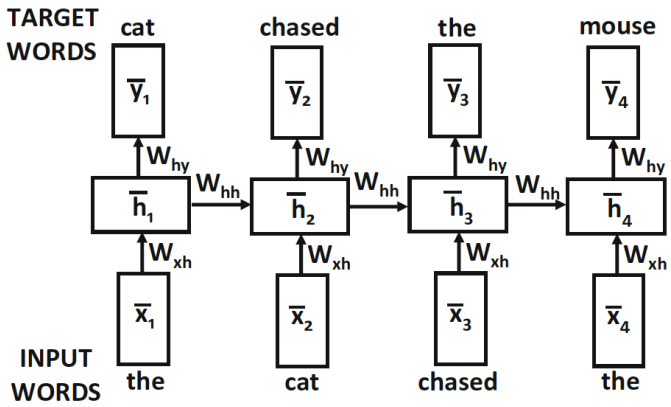


Fig. 3.8 Time-layered architecture of an RNN [5].

layers within the network and the specific timestamp or position in the sequence [5]. An RNN comprises a variable number of layers, with each layer having a single input corresponding to that particular timestamp. Specifically, considering a simple recurrent node in an RNN with the following equation:

$$h_t = \sigma(w_{xh} \cdot x_t + w_{hh} \cdot h_{t-1} + b_h) \tag{3.12}$$

where:

- $\boldsymbol{h_t}$: Hidden state at time t.
- $\boldsymbol{\sigma}$: Activation function applied element-wise.
- $\boldsymbol{w_{xh}}$: Weight matrix connecting the input $\boldsymbol{x_t}$ to the hidden state.
- $\boldsymbol{w_{hh}}$: Weight matrix connecting the previous hidden state $\boldsymbol{h_{t-1}}$ to the current hidden state.
- $\boldsymbol{b_h}$: Bias term for the hidden state.

RNNs can also be regarded as feed-forward networks with a specific structure rooted in the concept of time layering, enabling them to accept a sequence of inputs and generate a sequence of outputs. These models are proven to be particularly valuable for applications involving sequence-to-sequence learning, such as machine translation or predicting the next element in a sequence. In Fig. 3.8, the representation of an RNN is depicted. This architecture allows for the distinction of:

- **Input Sequence ($\boldsymbol{x_t}$):** Represents the input at time t, varying from the sequence's start to end.

- **Hidden State (h_t):** Denotes the hidden state at time t, capturing memory from previous steps.
- **Weights (w):** Learned parameters, including connections from input to hidden states and recurrent connections.
- **Output (y_t):** Represents the output at time t, predicting or representing the input sequence.

Long Short-Term Memory (LSTM) RNNs, in particular, excel at efficient parameter sharing through gated memory mechanisms [205]. Within each LSTM cell, recurrent units equipped with self-learned gating enable the preservation, modification, and selective forgetting of information within a short-term memory [206]. This effectively addresses challenges related to handling long learning dependencies, which can be problematic in other RNN variants.

3.5 Machine Learning Combined with Atomistic Simulations

Atomistic simulations are a key tool for exploring material mechanics. The fidelity of simulation results relies on the interatomic potential describing atom interactions. Classical potentials have two main limitations: transferability and version-control of the originally developed potentials [207]. The first is due to fixed forms and few fitting parameters. The second major issue is the risk of discrepancies between the implemented potential and the original version provided by developers. Maintaining accurate parameters over time is difficult due to file format changes, transfer errors, and file corruption. In contrast, MLIPs [87] offer flexibility by learning from first principles calculations rather than relying on fixed forms [88–90]. Successful ML potentials have been developed for various materials. MLIPs can be broadly categorized into two main types: Descriptor-based MLIPs use descriptors to characterize atomic environments, ensuring necessary invariances and uniqueness [208]. On the other hand, graph-based MLIPs learn atom environments directly from types and positions without a fixed descriptor, explicitly incorporating many-body interactions [209]. In this thesis, a descriptor-based approach will be employed. In computational materials science, especially with databases already mentioned in Sec. 2.1, ML potentials play a crucial role in swiftly identifying materials with desired properties. Representative approaches include [91, 92]:

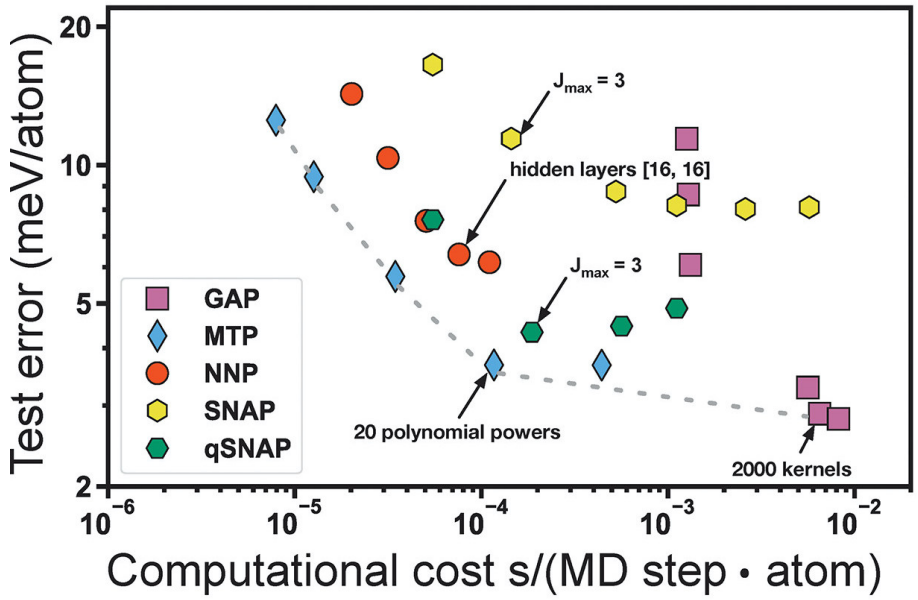


Fig. 3.9 Performance comparison of several descriptor-based ML potentials on a range of crystal structures [8].

- Behler and Parrinello Neural Network Potentials (NNP) [87]: Utilizes atom-centered symmetry functions (ACSF) and a high-dimensional neural network to describe the atom's local environment through radial and angular distribution functions.
- Spectral Neighbor Analysis Potentials (SNAP) [210]: Employs bispectrum components as the descriptor and a linear regression model to fit the data, representing another function expansion of local atomic density.
- Gaussian Approximation Potentials (GAP) [94]: Utilizes Smooth Overlap of Atomic Positions (SOAP) as the descriptor and Gaussian process regression as the ML model, employing SOAP as a function expansion of local atomic density.
- Moment Tensor Potentials (MTP) [95]: Uses rotationally covariant tensors along with a linear regression model, treating tensors as a series of radial and distribution functions similar to ACSF.

The performance of a ML potential depends on both the choice of descriptor and ML model. In Fig. 3.9, the performance comparison of descriptor-based ML potentials on a range of crystal structures with a single CPU core is depicted [8]. In light of the results from Fig. 3.9, the MTP, NNP, and SNAP models demonstrate significantly

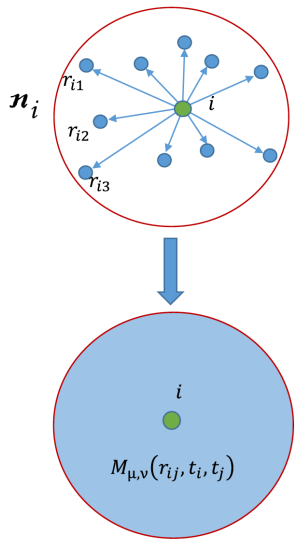


Fig. 3.10 Dimensionality reduction of the atomic neighborhood n_i for the atom i, described by the moment tensors $M_{\mu,\nu}$ [9].

higher computational efficiency than the GAP model. Nevertheless, achieving better accuracy necessitates a higher computational cost.

A well-trained ML potential, encompassing physics from diverse datasets, accurately predicts like first principles. Physics is integrated through PDML descriptors, acting as physico-chemical fingerprints rooted in fundamental laws and symmetry principles [211]. PDML uniquely encodes static chemical system details (nuclear charges and positions) governing the electronic Schrödinger equation. Notably, PDML's inherent transferability adapts to various tasks without task-specific tweaks, grounded in foundational laws, ensuring versatility across scenarios.

Descriptor-based MLIPs: Moment Tensor Potentials (MTPs)

Assuming that the potential energy E of an atomistic system can be well approximated by a sum of energies of atomic environments for individual atoms, denoted as D_{x_k}, the equation can be expressed as:

$$E(\mathbf{x}) = \sum_{k=1}^{n} V(D_{x_k}) \tag{3.13}$$

where D_{x_k} represents the atomic environment D_{x_k} for the k-th atom and $V(D_{x_k})$ the interatomic potential. An efficient alternative to V is MTPs, approximating the potential through invariant polynomials representing moments of the inertia tensor.

V is linearly expanded using basis functions dependent on moment tensors M. In Fig. 3.10, this descriptor is illustrated, which maintains invariance to permutations, rotations, and reflections, ensuring robust representation of interatomic interactions. This approach allows MLIPs with a linear dependence on fitting parameters, detailed in [95]. MTP derivation is implemented in the MLIP package [212].

Chapter 4

Simulation Methods

This chapter explores essential simulation techniques in molecular sciences, with a focus on DFT and Molecular Dynamics (MD), specifically addressing aspects relevant to this work. It mostly relies on references [213–215] for DFT and [216–218] for MD, providing a comprehensive overview. In DFT, it covers the time-independent Schrödinger equation, Hamiltonian, Born-Oppenheimer approximation, and Kohn-Sham equations, emphasizing the importance of the effective potential and exchange-correlation functionals. Transitioning to MD, it discusses Newton's equation of motion, integration and the application of thermostats, all of which are crucial for understanding the dynamics inherent in molecular systems.

4.1 Density Functional Theory

There are very few cases in which the Schrödinger equation can be solved analytically. Except for the Hydrogen atom, a well-known two-body problem, the study of multi-electron systems requires the use of numerical calculations and various approximations to simplify the Hamiltonian. Even in the simplest case, such as Helium with N = 2 electrons, analytical resolution is not possible. This challenge prompted the development of the two methods examined in this work: the Hartree-Fock model and the Kohn-Sham approximation within DFT. The goal in computational material science is to solve the complete time-independent many-body Schrödinger equation. Before delving into the details of DFT, it is beneficial to introduce the time-independent Schrödinger equation for the many-body problem. This equation encompasses the Hamiltonian of a multi-electronic system, along with the underlying hypotheses and general approximations explained below:

$$\hat{H}\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{4.1}$$

where:

- $\hat{H}$ represents the Hamiltonian operator, which describes the total energy of the system given by:

$$\hat{H} = [\hat{T} + \hat{V}] = [\hat{T}^e + \hat{T}^n + \hat{V}^{ee} + \hat{V}^{en} + \hat{V}^{nn}] \tag{4.2}$$

  where the indices e and n correspond to electrons and nuclei, and their combinations represent interactions between them:

  - $\hat{T}^e$: Kinetic energy operator for electrons.
  - $\hat{T}^n$: Kinetic energy operator for nuclei.
  - $\hat{V}^{ee}$: Potential energy operator for electron-electron repulsion.
  - $\hat{V}^{en}$: Potential energy operator for electron-nucleus attraction.
  - $\hat{V}^{nn}$: Potential energy operator for nucleus-nucleus repulsion.

- $\psi(\mathbf{r})$ is the wavefunction, representing the quantum state of the system as a function of position $\mathbf{r}$.

- E is the energy eigenvalue associated with the wavefunction $\psi(\mathbf{r})$.

Concerning Eq. 4.2, the potential of atoms or molecules is determined through non-empirical methods, like *ab initio* calculations, grounded in quantum physics and the Schrödinger equation. These calculations typically involve systems with atoms, nuclei, and electrons. The primary challenge in analytical resolution is the potential terms, given the wavefunction's dependence on both nuclear and electronic coordinates, which are non-separable. Here is where general approximations come into play:

- **Born-Oppenheimer (BO) Approximation:** Simplifies the potential term by considering the mass difference between nuclei and electrons. Since nuclei are about 2000 times heavier[1], they can be considered static compared to electrons, leading to the elimination of the nuclear kinetic term $\hat{T}_n$ and the nucleus-nucleus interaction $\hat{V}_{nn}$ from the potential, which is treated as constant.

$$\hat{H}_{\mathrm{BO}}\psi(\mathbf{r}, \mathbf{R}) = E_{\mathrm{elec}}(\mathbf{R})\psi(\mathbf{r}, \mathbf{R}) \tag{4.3}$$

  where:

  - $\hat{H}_{\text{BO}}$: The BO electronic Hamiltonian operator.
  - $\psi(\mathbf{r}, \mathbf{R})$: The total wavefunction, dependent on both electronic coordinates $\mathbf{r}$ and nuclear coordinates $\mathbf{R}$.
  - $E_{\text{elec}}(\mathbf{R})$: The electronic energy, parametrically dependent on nuclear coordinates $\mathbf{R}$ but independent of electronic coordinates $\mathbf{r}$.

- **Independent Electron Approximation:** Even when applying the BO Hamiltonian, the problem for electronic wavefunctions remains unsolvable due to the number of degrees of freedom involved. Hence, this approximation decouples the system, modeling the electron-electron interaction term as the sum of independent electronic Hamiltonians.

Additionally, both methods must be self-consistent. This is crucial to ensure convergence to the ground state when iteratively solving the Schrödinger equation. Considering these approximations, the BO Hamiltonian in atomic units can be reformulated as:

$$\hat{H}_{BO}\psi = [\hat{T}^e + \hat{V}^{en} + \hat{V}^{ee}]\psi = [-\sum_{i=1}^{N_e} \frac{1}{2}\nabla_i^2 + \sum_{i}^{N_n}\sum_{j}^{N_e} \frac{Z_i}{|\vec{r_j} - \vec{R_i}|} + \sum_{i<j}^{N} V(\vec{r_i}, \vec{r_j})]\psi = E\psi \tag{4.4}$$

where:

- $\hat{T}^e$: Kinetic energy operator for electrons, represented by $-\sum_i^N \frac{1}{2}\nabla_i^2$, where i ranges over all electrons (N in total).
- $\hat{V}^{en}$: Potential energy operator for electron-nucleus attraction, given by $\sum_n^N \frac{Z_n}{|r-R_n|}$, where n ranges over all nuclei (N in total), Z_n is the charge of the n-th nucleus, r is the electron's position, and R_n is the position of the n-th nucleus.
- $\hat{V}^{ee}$: Potential energy operator for electron-electron repulsion, expressed as $\sum_{i<j}^N V(\vec{r_i}, \vec{r_j})$, where i and j range over all distinct pairs of electrons (N in total), and $V(\vec{r_i}, \vec{r_j})$ is the interaction potential between electrons at positions $\vec{r_i}$ and $\vec{r_j}$.

4.1.1 Kohn-Sham Equations

The Kohn-Sham (KS) DFT considers a fictitious system of non-interacting electrons that mimics the electron density of the actual system. To simplify the problem, the true many-electron wavefunction Ψ is approximated using a set of single-electron wavefunctions $\psi_i(\mathbf{r})$, known as KS orbitals. KS DFT is grounded in the concept of electron

density $n(\mathbf{r})$ as defined in quantum theory, aiming to simplify the electron-electron potential. This non-interacting system is described by single-electron wavefunctions $\psi_i(\mathbf{r})$ with the same electron density as $n(\mathbf{r})$. The electron density is defined as the sum of the squared magnitudes of the KS orbitals:

$$n(\mathbf{r}) = \sum_{i}^{N} |\psi_i(\mathbf{r})|^2 \tag{4.5}$$

Here,

- $n(\mathbf{r})$: Electronic density,
- N: Number of electrons, and
- $\psi_i(\mathbf{r})$: Kohn-Sham orbitals.

The KS DFT framework relies on the two theorems of Hohenberg-Kohn:

- **First Theorem (Uniqueness):** The ground-state electronic density $n(\mathbf{r})$ uniquely determines the ground-state wavefunction and, consequently, all other ground-state properties, including the energy E_G.

$$E[n(\mathbf{r})] = F[n(\mathbf{r})] + \int V_{\text{ext}}(\mathbf{r}) n(\mathbf{r})\, d\mathbf{r} \tag{4.6}$$

  where $F[\rho(\mathbf{r})]$ is a universal functional of density.

- **Second Theorem (Existence):** There exists a one-to-one correspondence between the external potential $V_{\text{ext}}(\mathbf{r})$ and the ground-state electron density $n(\mathbf{r})$. In other words, different external potentials cannot lead to the same ground-state density.

In essence, these theorems establish a direct link between electron density and ground state energy, facilitated by the existence of a unique density functional. This enables precise calculations without the need for wavefunctions of all electrons. The challenge lies in determining the universal functional governing the exact ground state energy. Attempts to approximate, like the Thomas-Fermi model [219, 220], fell short in predicting kinetic energy contributions accurately. The issue found resolution in the Kohn and Sahm approximation, introducing a correlation-exchange potential V_{XC} as the functional derivative of the correlation-exchange energy E_{XC}. This term signifies the kinetic energy of a non-interacting electron cloud, upholding the electronic independence

of BO. Manipulating this term in the DFT Hamiltonian aims to consolidate the entire unknown part of the functional independent of the external potential. The KS electrons do not interact with each other but rather interact with an external potential V_{XC}. This interaction is designed such that their ground-state charge density precisely matches that of the interacting system. Derived from these theorems, the total energy can be expressed as a functional of the electron density:

$$E[n] = T_s[n(\mathbf{r})] + \int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})\, d\mathbf{r} + \frac{1}{2}\int d\mathbf{r}d\mathbf{r}'\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\text{xc}}[n(\mathbf{r})] \tag{4.7}$$

where:

- $E[n]$: Total electronic energy functional of the electron density $n(r)$.
- $T_s[n]$: Kinetic energy of non-interacting electrons.
- $\int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})\, d\mathbf{r}$: Electron-nuclear attraction energy.
- $\frac{1}{2}\int drdr'\frac{n(r)n(r')}{|r-r'|}$: Electron-electron repulsion energy.
- $E_{\text{xc}}[n]$: Exchange-correlation energy.

The KS equations, derived through variational minimization using the Lagrange multipliers method, describe non-interacting electrons in an effective potential replicating the real system's electron density. This approach helps identify the most stable electron density, forming the foundation for the KS equations. These equations play a crucial role in density functional theory, offering a potent tool for exploring the electronic properties of complex systems. Now, we apply the Lagrange multipliers method to enforce the orthonormality of the single one-particle wavefunctions.

$$\frac{\delta[E[n] - \lambda_i \int d\mathbf{r}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}) - 1]}{\delta\psi_i^*} = 0 \tag{4.8}$$

In this unit-less equation, $\frac{\delta}{\delta\psi_i^*}$ represents the functional derivative with respect to the complex conjugate of the i-th single-electron wavefunction $\psi_i(\mathbf{r})$, $E[n]$ is the total energy functional of the electron density $n(\mathbf{r})$, and λ_i is a Lagrange multiplier associated with the normalization constraint of the wavefunction $\psi_i(\mathbf{r})$. The equation is set to zero to find the stationary point and solve for the KS orbitals in the self-consistent procedure. For a comprehensive understanding of the derivation process for the KS equations, please refer to the details provided in the references [10, 221, 222]. Concluding with the previous approximations, the KS DFT Hamiltonian of the system is given by:

$$\hat{H}_{\text{KS DFT}} = -\frac{1}{2}\nabla_i^2 + V_{\text{eff}}(r) = -\frac{1}{2}\nabla_i^2 + V_{\text{ext}}(r) + \int dr' \frac{n(r')}{|r - r'|} + V_{\text{XC}}(r) \tag{4.9}$$

where:

- $-\frac{1}{2}\nabla_i^2$: Kinetic energy operator for electrons.
- $V_{\text{ext}}(r)$: External potential due to electron-nucleus interaction.
- $\int dr' \frac{n(r')}{|r-r'|}$: Also Hartree term $V_{\text{Hartree}}(r)$, representing electron-electron repulsion.
- $V_{\text{XC}}(r)$: Exchange-correlation potential, incorporating effects beyond Hartree.

The effective potential, denoted as $V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})$, plays a crucial role in KS DFT. Since $\hat{V}_{\text{eff}}$ depends on the KS orbitals, solving the KS DFT Schrödinger equation requires iterative self-consistent solutions until satisfactory convergence is achieved. The density of the fully interacting system can be precisely determined from V_{eff} by solving N one-electron equations. This process significantly simplifies the fully interacting problem. The KS DFT Schrödinger equation is expressed as follows:

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}) \tag{4.10}$$

The eigenvalues of the KS equation, denoted as ϵ_i, can be accurately calculated using numerical techniques that involve employing an appropriate E_{XC} functional in the self-consistent calculations. The ground-state energy of the system is determined by iteratively calculating $V_{\text{eff}}(\mathbf{r})$, updating the KS orbitals $\psi_i(\mathbf{r})$, and recalculating the density $n(\mathbf{r})$ from the resulting orbitals. This iterative procedure continues until self-consistency between the potential and electron density is achieved.

4.1.2 Exchange-correlation Functionals

The development of DFT over the past half-century, since the introduction of the KS method, has aimed at accurately approximating the exact exchange and correlation energy functional. The goal is to enhance DFT functionals for obtaining precise energies and electronic properties. The exchange-correlation potential $V_{\text{XC}}(r)$ term can be approximated using various methods, including the local density approximation (LDA), the generalized gradient approximation (GGA), and their mixture known as hybrid functionals, among others. According to [223], these $V_{\text{XC}}(r)$ functionals can be organized on a 'Jacob's ladder.' When no exchange-correlation functional is used, KS

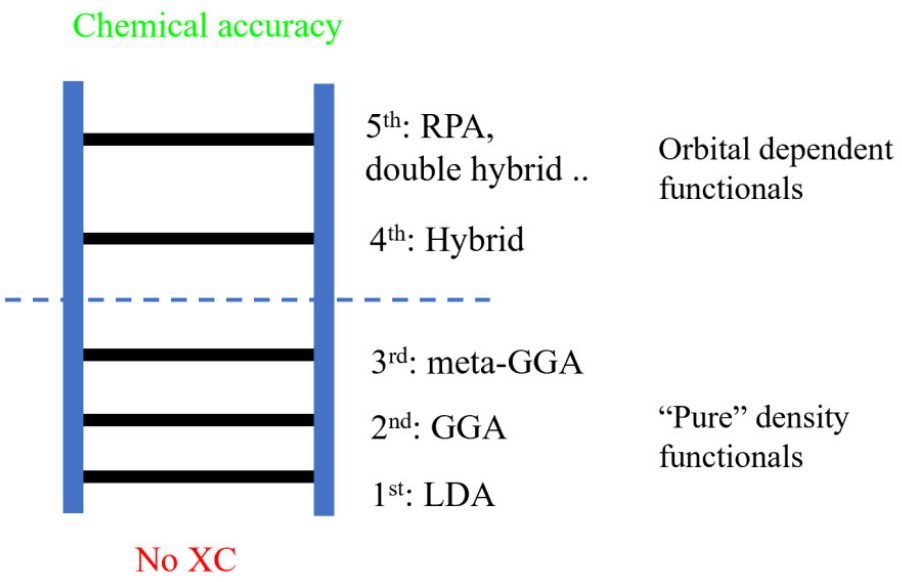


Fig. 4.1 The 'Jacob's ladder' of exchange-correlation functionals [10].

DFT essentially becomes a Hartree approximation. In Fig. 4.1, the Jacob's ladder is depicted. As one moves up the ladder, the accuracy of the DFT calculation generally improves. The formulation of LDA and GGA functionals can be written as:

- **LDA (Local Density Approximation):**

$$E_{\mathrm{xc,LDA}}[n] = \int n(\mathbf{r})\varepsilon_{\mathrm{xc,LDA}}(n(\mathbf{r}))\, d\mathbf{r} \tag{4.11}$$

- **GGA (Generalized Gradient Approximation):**

$$E_{\mathrm{xc,GGA}}[n] = \int n(\mathbf{r})\varepsilon_{\mathrm{xc,GGA}}(n(\mathbf{r}), \nabla n(\mathbf{r}))\, d\mathbf{r} \tag{4.12}$$

These equations represent the exchange-correlation energy functionals for LDA and GGA, where $n(\mathbf{r})$ is the electron density, and $\varepsilon_{\mathrm{xc}}$ denotes the exchange-correlation energy per particle. The GGA functional explicitly depends on the gradient of the electron density, denoted as $\nabla n(\mathbf{r})$. LDA is employed to simplify the correlation-exchange energy term. This approximation assumes that the system is homogeneous enough to exhibit behavior closely resembling that of an electron cloud, where the density varies slowly at each point (i.e., is quasi-homogeneous). The energy expression for this system is known as a functional, enabling the iterative resolution of the KS DFT equations. GGA introduces a correction to the gradient of the ground state

electron density, represented as $\nabla n(\mathbf{r})$. Although LDA frequently produces satisfactory outcomes, the preference for GGA or LDA varies with the system in question. There is no universal assurance that GGA consistently surpasses LDA; the decision often revolves around achieving the optimal trade-off between computational accuracy and costs.

In this thesis, we specifically use a sub-class of the GGA functional developed by Perdew, Burke, and Ernzerhof (PBE), GGA PBE exchange-correlation functional:

$$E_{\mathrm{xc,PBE}}[n] = \int n(\mathbf{r})\varepsilon_{\mathrm{xc,PBE}}(n(\mathbf{r}), \nabla n(\mathbf{r}))\, d\mathbf{r} \tag{4.13}$$

The GGA PBE functional refines the description by accurately representing the linear response of the uniform electron gas, ensuring correct behavior under uniform scaling, and yielding a smoother potential. It retains the essential features of LDA while incorporating crucial gradient-corrected nonlocality. The PBE exchange term ($\varepsilon_{\mathrm{x}}^{\mathrm{PBE}}$) includes an enhancement factor for a more accurate depiction of electrons in a uniform electron gas. The correlation term ($\varepsilon_{\mathrm{c}}^{\mathrm{PBE}}$) combines the LDA correlation term with a correction term incorporating gradient corrections. This correction is a logarithmic function ($\Delta\varepsilon_{\mathrm{c}}^{\mathrm{PBE}}$) accounting for gradient corrections. The specificity of the PBE lies in these tailored exchange and correlation functionals, enhancing accuracy in describing electron behavior at varying densities and gradients. For a more detailed overview of the GGA PBE functional, please refer to [224].

4.1.3 Crystals: Periodicity, Basis Sets and Pseudopotentials

In Chapter 5 of this thesis, the analysis of HEAs with a crystalline structure is explored, simplifying them into unit cells for computational efficiency. Utilizing periodic boundary conditions, the representation of the KS orbitals efficiently in a plane-wave basis set following Bloch's theorem is undertaken [225]:

$$\psi_k(\mathbf{r}) = \sum_{\mathbf{G}} c_{k+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \tag{4.14}$$

where $\mathbf{G}$ is a reciprocal lattice vector, and $\mathbf{k}$ is a vector in reciprocal space within the Brillouin zone. The plane-wave basis size is determined by an energy cutoff for the kinetic energy:

$$\frac{\hbar^2}{2m_e}|\mathbf{k}+\mathbf{G}|^2 < E_{\mathrm{cut}}. \tag{4.15}$$

Achieving reliable DFT results demands increasing (E_{cut}) until total energy convergence. Valence KS orbitals, exhibiting oscillations in the core region due to orthogonality requirements with core KS orbitals, necessitate a large energy cutoff, resulting in higher computational costs. In order to address this, pseudopotentials are employed, which freeze core electrons, consolidate nuclei and core-electron potentials into a smooth potential, and reduce KS equation solving costs by lowering required orbitals and permitting lower energy cutoffs. This minimizes oscillations in valence KS orbitals. The pseudopotentials used in the DFT calculations of this thesis are derived from all-electron atomic calculations, solving the radial KS equation self-consistently.

4.2 Molecular Dynamics

MD is a prevalent simulation method employed to examine the dynamics of classical many-particle systems at the molecular scale. The approach represents particles as point masses, sometimes with associated charges. MD employs collections of molecular models and parameterized interaction potentials, also known as force fields, whose parameters depend on the selected molecular models. The MD approach is not only proficient in examining static physical parameters but also excels in exploring dynamic properties far from equilibrium, such as diffusion coefficients and ionic conductivity. During MD simulations, particles adhere to Newton's second law, $\mathbf{F}_i = m_i \mathbf{a}_i$, with forces derived from the potential $V(\mathbf{r}_i, t)$ derivative concerning the particle's position $\mathbf{r}_i$, thereby defining the equation of motion for each particle.

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} = -\frac{dV(\mathbf{r}_i, t)}{d\mathbf{r}_i} \tag{4.16}$$

where:

- m_i is the mass of particle i.
- $\mathbf{r}_i$ is the position vector of particle i in three-dimensional space as a function of time t.
- $\frac{d^2\mathbf{r}_i}{dt^2}$ represents the second derivative of $\mathbf{r}_i$ with respect to time t, which corresponds to the acceleration $\mathbf{a}_i$ of particle i.
- $\mathbf{F}_i$ is the net force acting on particle i due to interactions with other particles or external fields.

- V represents the potential, referred to as the force field in MD. Its dependence extends to the collective positions of all particles $\mathbf{r}_i$ and, in instances involving time-dependent external fields, also on time t. The determination of V can take an empirical approach or be derived through *ab initio* calculations.

Classical MD simulations utilize approximations, treating particle forces as classical and conservative, described by distance-dependent effective potentials or force fields. The solution of the equations of motion is approximate for the following reasons:

- All-atom MD simulations model atoms, and their interactions follow quantum mechanics principles. However, evaluating many-body interactions quantum-mechanically for moderately sized molecular systems is computationally challenging.

- The forces between particles are not known exactly. The interaction between particles is calculated using predefined interaction parameters corresponding to a certain force-field.

In this thesis, ~~our~~ MD simulations will employ the MEAM potential, complemented by ML interatomic potentials derived from prior *ab initio* calculations. The Modified Embedded Atom Method (MEAM) [226, 227] is a molecular dynamics potential that employs a semi-empirical approach to model atomic interactions. It is an extension of the Embedded Atom Method (EAM) [228, 229] with refinements for enhanced accuracy. The MEAM potential comprises equations governing embedding energy, pair interactions, and triplet interactions. A representative equation is as follows:

$$E = \sum_i F(\rho_i) + \frac{1}{2}\sum_{i,j} \phi(\rho_i, \rho_j, \vec{r}_{ij}) + \frac{1}{3}\sum_{i,j,k} \psi(\rho_i, \rho_j, \rho_k, \vec{r}_{ij}, \vec{r}_{ik}, \vec{r}_{jk}) \tag{4.17}$$

Here, ρ_i is electron density at atom i, and $\vec{r}_{ij}$ is the distance between atoms i and j. The functions F, ϕ, and ψ capture embedding, pair, and triplet interactions, respectively. The specific forms depend on the chosen MEAM potential for the material.

4.2.1 Integraton Schemes

In order to integrate Newton's equations of motion in MD, numerical integration methods are employed for propagating the positions and velocities of particles over time. One widely utilized numerical integration method in MD that yields the positions, velocities and forces at the same time is the Verlet algorithm, which serves as the

default scheme in LAMMPS [230] and is commonly known as the Störmer-Verlet time integration algorithm or simply Verlet integration [231]. In this thesis, the Velocity-Verlet algorithm, the most commonly used algorithm in practice, is utilized to integrate the positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ of particles i over a time step Δt:

$$\mathbf{v}_i(t + \frac{\Delta t}{2}) = \mathbf{v}_i(t) + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t \tag{4.18}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{\Delta t}{2})\Delta t \tag{4.19}$$

$$\mathbf{F}_i(t + \Delta t) = \mathbf{F}(\mathbf{r}_i(t + \Delta t)) \tag{4.20}$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t + \frac{\Delta t}{2}) + \frac{\mathbf{F}_i(t + \Delta t)}{2m_i}\Delta t \tag{4.21}$$

where:

- $\mathbf{r}_i(t)$ is the position vector of particle i at time t.
- $\mathbf{v}_i(t)$ is the velocity vector of particle i at time t.
- $\mathbf{F}_i(t)$ is the net force acting on particle i at time t due to interactions with other particles or external fields.
- m_i is the mass of particle i.
- Δt is the time step used in the integration, which determines the granularity of the simulation.

By repeatedly applying the steps from Eq. 4.18 to 4.21, the positions and velocities of all particles can be propagated over time, simulating the dynamics of the molecular system in molecular dynamics simulations. The Verlet algorithm is computationally efficient and widely used due to its simplicity and stability.

4.3 Statistical Ensembles in MD

In statistical physics, MD simulations explore the temporal evolution of microscopic configurations representing possible microstates in an equivalent thermodynamic system. The collective states form a statistical ensemble, with corresponding ensembles depending on the defining thermodynamic variables of the system. Various statistical ensembles offer distinct perspectives on molecular state distribution, aiding in thermodynamic property analysis:

- **Microcanonical Ensemble:** The Microcanonical Ensemble, or NVE (Number of particles, Volume, Energy) ensemble, isolates a system from external heat baths, keeping the total energy constant. This ensemble is particularly useful for studying isolated systems where energy conservation is crucial.

- **Canonical Ensemble:** The Canonical Ensemble, also known as the NVT (Number of particles, Volume, Temperature) ensemble, is commonly used in MD simulations. In this ensemble, the system is in thermal contact with a heat bath, maintaining a constant temperature. It allows for the exploration of equilibrium properties such as temperature-dependent fluctuations.

- **Isothermal-isobaric Ensemble (NpT):** The Isothermal-isobaric Ensemble (NpT) describes a system with a fixed number of particles (N), allowing fluctuations in both pressure (p) and temperature (T). This ensemble models systems in contact with a heat bath to maintain a constant temperature and in contact with a reservoir to maintain a constant pressure. It is suitable for simulating realistic conditions where temperature and pressure play crucial roles, such as in chemical reactions and phase transitions.

- **Grand Canonical Ensemble:** For simulations involving systems with the exchange of particles with a reservoir, the Grand Canonical Ensemble (μVT) is employed. This ensemble maintains a constant chemical potential, enabling the study of systems under various chemical conditions.

Each ensemble provides a unique perspective on the molecular dynamics of a system, offering a comprehensive understanding of its thermodynamic behavior.

4.4 Thermostats

Thermostats act like algorithms designed to control and maintain the temperature of a simulated system by regulating the velocities of particles. This ensures that kinetic energy remains consistent and aligns with the desired temperature. Thermostats play a crucial role in emulating realistic thermal behavior, allowing simulations to explore equilibrium properties under specific temperature conditions. The earliest method for constant temperature MD involves a momentum scaling procedure, where particle velocities are scaled at each time step to maintain the total kinetic energy at a constant value. Other popular thermostats include the Berendsen thermostat, Andersen thermostat, and the Nosé-Hoover thermostat [232, 233]—the latter being utilized in

this thesis. Each offers distinct approaches to temperature control in MD simulations. The standard Nosé-Hoover thermostat is known to have issues, particularly for small or stiff systems, where the dynamics may not be ergodic. These can be addressed by using multiple Nosé-Hoover thermostats connected in a chain M [234]. The inclusion of the chain of thermostats M in the extended Nosé-Hoover thermostat enhances its performance, providing more accurate and stable temperature control. The equations of motion for the Nosé-Hoover chain method are:

$$\frac{dr_i}{dt} = \frac{p_i}{m_i} \tag{4.22}$$

$$\frac{dp_i}{dt} = F_i - p_i \frac{p_{\eta_1}}{Q_1} \tag{4.23}$$

$$\frac{d\eta_i}{dt} = \frac{p_{\eta_i}}{Q_i} \tag{4.24}$$

$$\frac{dp_{\eta_1}}{dt} = \left[\sum_{i=1}^{N} \frac{p_i^2}{m_i} - N_f kT\right] - p_{\eta_1} \frac{p_{\eta_2}}{Q_2} \tag{4.25}$$

$$\frac{dp_{\eta_j}}{dt} = \left[\frac{p_{\eta_{j-1}}^2}{Q_{j-1}} - kT\right] - p_{\eta_j} \frac{p_{\eta_{j+1}}}{Q_{j+1}} \tag{4.26}$$

$$\frac{dp_{\eta_M}}{dt} = \left[\frac{p_{\eta_{M-1}}^2}{2Q_{M-1}} - kT\right] \tag{4.27}$$

These equations govern the positions (r_i) and momenta (p_i) of atoms, responding to forces (F_i). These variables interact with auxiliary (η_i), included for completeness, and thermostat momenta (p_{η_i}), key for temperature control. The Q_i terms denote thermostat mass, influencing inertia. Friction terms ($p_i \frac{p_{\eta_1}}{Q_1}$, $p_{\eta_j} \frac{p_{\eta_{j+1}}}{Q_{j+1}}$) capture molecular-thermostat coupling. Expressions like $\sum_{i=1}^{N} \frac{p_i^2}{m_i}$ and NkT manage energy and temperature. This interplay, involving M thermostats, η, p_η, Q, k, and N_f (degrees of freedom), shapes the dynamics for accurate simulations. The enhanced Nosé-Hoover thermostat, utilized as an NVT-ensemble integrator, incorporates the Verlet algorithm alongside a Nosé-Hoover chain thermostat with a chain length of M. For more in-depth information about the integrator, readers can consult [234].

Chapter 5

Stability of n-ary Phases in Cu-Ni-Si-Cr Alloys

This study employs computer simulations and ML to investigate complex phases in copper alloys. By applying AL and generating MTPs based on QM calculations, the research explores novel candidates and assesses their stability. Focusing on Cu-Ni-Si-Cr alloys, binary complexes such as Cu-Si, Ni-Si, Cr-Si, Cr-Ni, Cu-Ni, Cu-Cr, and the quaternary Cu-Ni-Si-Cr systems are analyzed for stability. The approach successfully predicts binary phases and extends known structures in the AFLOW library. The study concludes by demonstrating the application of predicted phases for calculating mechanical properties and melting behavior, presenting an efficient framework for material structure prediction and property calculation at QM accuracy.

5.1 Computational Details

5.1.1 Crystal Prototype Sampling

The process of generating crystal prototypes for materials discovery is a challenging task, primarily due to the vast phase space involved and the high computational cost of first-principle calculations. To expedite the prediction of new materials, this study utilizes the ENUMLIB library [235]. This library offers an algorithm that generates derivative superstructures from systems with fcc, bcc, and hcp symmetries, effectively enumerating various atomic configurations and super-lattices in a geometry-independent manner. These prototypes play a crucial role in exhaustive searches of binary configurations on lattices, enabling the determination of ground state properties in intermetallic systems [236]. Furthermore, the library incorporates Vegard's law to

System	fcc	bcc	AFLOW	Other	Total
Cu-Si	10850	12858	1	3	23712
Ni-Si	10850	12858	5	1	23714
Cr-Si	10850	12858	2	3	23713
Cr-Ni	10850	12858	1	1	23710
Cu-Ni	10850	12858	2	0	23710
Cu-Cr	10850	12858	0	0	23708
Cu-Ni-Si-Cr	4158	-	0	1	4159

Table 5.1 Number of prototypes for the binary and quaternary systems studied in this work, generated by ENUMLIB for the parent lattices fcc and bcc. 'Other' refers to the number of relevant structures taken from one or more of the ICSD, OQMD and Materials Project databases.

estimate the initial lattice constant for each prototype [237, 238]. In this work, the procedure is twofold:

- For the binary prototypes, the fcc and bcc symmetries serve as parent lattices for the input of the ENUMLIB generator. The atoms are subsequently labeled with the respective species under investigation. These two sets of generated candidates are referred to as the 'fcc set' and 'bcc set'. A total of 23,708 prototypes are generated for each binary system, namely Cu-Si, Ni-Si, Cr-Si, Cr-Ni, Cu-Ni, and Cu-Cr, as outlined in Tab. 5.1. These structures are then employed as inputs for the AL relaxation on-the-fly algorithm. Additionally, experimentally observed structures from repositories such as AFLOW [79, 239], ICSD, OQMD, and Materials Project are also included in the analysis.

- For the quaternary prototypes, the number of unique stoichiometries and their respective candidates, generated from an fcc parent lattice, are shown in Tab. 5.2. In total, 4,159 prototypes are generated for the quaternary system. These structures are used as inputs for the AL relaxation on-the-fly algorithm. In addition and already included on those 4,159, an fcc structure reported in the repository OQMD was added to the set.

5.1.2 On-the-fly Active Learning Relaxation: AL-MTP

Static QM simulations performed at zero temperature are combined with AL. This allows the configurational space to be sampled during simulation, in order to reveal stable precipitate phases — in this case — of copper alloy relevant binaries. This method, referred to as 'AL-MTP', is implemented in the software package MLIP [212].

Formula	Cell size	Samples
CuSiNiCr	4	19
$Cu_2NiSiCr$	5	27
$Cu_2Ni_2Si_2Cr_2$	8	2404
$Cu_3NiSiCr$	6	100
$Cu_3Ni_2Si_2Cr$	8	1571
$Cu_4NiSiCr_2$	8	8
Cu_4NiSi_2Cr	8	16
Cu_4Ni_2SiCr	8	8
$Cu_5NiSiCr$	8	4

Table 5.2 Number of unique stoichiometries for the quaternary systems studied in this work, generated by ENUMLIB for the parent lattice fcc.

The basis of this method are the MTPs, a class of descriptors based on invariant polynomials that has successfully been used in the case of multi-component alloys. The iterative AL-MTP algorithm works as follows [9]: The input to the algorithm is a set of structure candidates to be relaxed, an MTP functional form $E = E(\Theta, x)$ determined by the level lev_{max} (with Θ the MTP parameters), outputs from QM simulations and two thresholds γ_{tsh} and Γ_{tsh}, such that $\Gamma_{\text{tsh}} > \gamma_{\text{tsh}} > 1$. The steps below briefly summarize the AL-MTP approach and are followed for the set of all structural candidates considered here and inferred in Tab 5.1:

(1) A structure relaxation is performed for each candidate with extrapolation grade $\gamma(x)$, where two scenarios are possible:

   – The candidate converges to an equilibrium state, where $\Gamma_{\text{tsh}} > \gamma(x)$ for each relaxation step. The structure for which at an intermediate step $\Gamma_{\text{tsh}} > \gamma(x) > \gamma_{\text{tsh}}$ is added to the pre-selected set.

   – The relaxation stops due to $\gamma(x) > \Gamma_{\text{tsh}}$. This situation indicates that the MTP extrapolation is too high. The last and previous configurations with $\gamma(x) > \gamma_{\text{tsh}}$ are added to the pre-selected set.

(2) Out of the pre-selected set created during step (1) a small number of relevant candidates is selected through the D-optimality criterion. This optimal design technique provides a measurement for the proximity of a certain structure to the training set of the MTP. After this and for each pre-selected structure a static QM calculation is performed in VASP. The corresponding energies, forces and stresses are added to the training set, to which we refer as 'pre-selected' data set in the following.

(3) The MTP is fitted to the updated training set.

(4) The steps 1-3 are repeated until all candidates have been successfully relaxed, according to the convergence criterion of the L-BFGS internal algorithm in MLIP. In this case the force tolerance is set to 0.0001 eV.

The range of extrapolation grade and level values used in the AL-MTP procedure for the relaxation of the binary precipitates of Cu-Ni-Si-Cr alloys in this work are $\gamma_{\text{tsh}} \in [2.0-5.0]$, $\Gamma_{\text{tsh}} \in [5.0-10.0]$, and $\text{lev}_{\text{max}} \in [16-20]$. The maximum cut-off radius of the MTP potential generated was set at 5Å. For the static QM calculations, the DFT implementation in VASP (version 5.4.4) was used. with the Perdew–Burke–Ernzerhof (PBE) [240] exchange-correlation functional and the projector-augmented wave (PAW) pseudopotentials [241]. The energy cutoff values were optimized in the range ∈ [430-520] eV for all the binary systems. The k-point mesh was generated automatically with a spacing between k-points of (kspacing) 0.15 Å^{-1}.

5.1.3 Convex Hull Calculations

The convex hull is considered as a mapping of the stability of multi-component systems based on the formation enthalpy throughout the concentration range of a certain element included in the system. For a certain structure, the formation enthalpy can be calculated using Hess' law:

$$\Delta H_f = \frac{1}{N}\left[E^{total} - \sum_{i}^{M} \frac{n_i}{N_i} \cdot E_i^{total}\right] \tag{5.1}$$

where N and E^{total} are the number of atoms and the MTP total energy for each prototype structure, respectively. n_i is the number of atoms in the unit cell for each species of the prototype and N_i is the number of atoms for the unit cell of each species in the respective one-component system. The term E^{total} is predicted via the AL-MTP algorithm, while the total energy E_i^{total} is calculated via volume relaxation in VASP. Finally, M is the number of species per structure. In this work, the value M=2 is used for the binary complexes, and M=4 is employed for the quaternary case. DFT calculations are performed for the unit cells of Cu, Ni, Si and Cr in order to find the respective ground-state energies, which correspond to the thermodynamic stable unit cell structures. For the relaxations, the VASP-recommended PAW PBE pseudopotentials Cu, Ni, Si and Cr_pv are used to carry out the DFT calculations. The magnetic behaviour of the alloying elements, such as non-magnetic (NM), ferromagnetic

(FM), or anti-ferromagnetic (AFM), has also been considered to add spin polarization whenever required. In Tab. 5.3, the details of the volume relaxation for each alloying element of the Cu-Ni-Si-Cr system are presented. The total energies and number of atoms from Tab 5.3 are implemented in Eq. 5.1 for the formation enthalpy calculations of the respective binary systems. In order to prove the reliability of the energy predictions and establish a direct comparison between our results and the AFLOW database, the vertices of the convex hulls found by the AL-MTP algorithm have been post-relaxed via DFT using the same input parameters. The latter also include the pseudopotentials given above for consistency and not the ones chosen in AFLOW.

i	symm	N_i	E_i^{total}	magn
Cu	fcc	4	-3.710	NM
Ni	fcc	4	-5.566	FM
Si	fcc	8	-5.423	NM
Cr	bcc	2	-9.631	AFM

Table 5.3 Ground state total energies (E_i^{total} in eV/atom) calculated during the DFT volume relaxation for the unit cells of the alloying elements i={Cu, Ni, Si, Cr} in the crystal symmetries ('symm') and magnetic state ('magn') stated. For the notation, see text and Eq. 5.1.

5.1.4 Calculation of Phonon Dispersion

In order to confirm the stability of the structures predicted through the AL-MTP approach and the convex hull calculation, the phonon modes for the minimum energy structures in the convex hulls of the binary systems are further analyzed. For this, the finite-displacement method implemented in Phonopy [242] was used. Positive phonon frequencies indicate that the curvature of the PES is also positive. On the other hand, imaginary phonon frequencies imply a negative curvature in the PES implying that the crystal does not correspond to its ground state energy or that relaxation has entered a local energy minimum [243]. Prior to the phonon modes calculation, all lowest energy structures predicted from the convex hull analysis were further post-relaxed. They were then replicated into a $3 \times 3 \times 3$ supercell with an atomic displacement of 0.01 Å. The corresponding static energy and atomic forces of the displaced configurations were calculated using the MD code LAMMPS [230] and the MTPs trained and generated in this work. The MTP potentials, referred to as 'MD-MLIP', use a cutoff radius of 5 Å. To ensure high accuracy in atomic force calculations, DFT phonon modes calculations were performed and compared to results obtained with MD-MLIP. Configurations with

the same displacement were tested within the DFT framework with a cutoff energy of 430 eV and a k-points mesh of $4 \times 4 \times 4$. The electronic self-consistent calculation end for energy differences below 1×10^{-7} eV [244]. In the end, the phonon modes were calculated directly for the DFT post-relaxed structures, as well as those obtained from the MD simulations and were compared.

5.1.5 Molecular Dynamics Simulations

In the end, MD simulations were carried out as implemented in the code LAMMPS with the LAMMPS-MLIP interface [230, 212]. The system was made of a bulk phase of Cu_7Si with a total of 512 atoms, with periodic boundary conditions used in all three directions of space. The initial atom positions were taken from a stable Cu_7Si relaxed unit cell as predicted through the AL-MTP procedure followed in this work. The elastic constants, the bulk modulus, and the shear modulus were calculated at zero temperature through the application of strain. The radial distribution function (RDF) between Cu and Si atoms, $g_{\text{Cu-Si}}(r)$, as well as the box volume V were calculated in the NPT ensemble with anisotropic cell fluctuations with a timestep of 1 fs. The pressure was set to $p = 1$ bar. For the analysis of the thermal response, the temperature T was increased step-wise from 300 K to 3000 K in 20 steps. For each of these steps, an equilibration of 20 ps was performed, followed by a production run of 200 ps. These simulations could have been run for longer durations but, for a solid material, long time-scales are typically not involved, as there are no slow processes. Therefore, extending the simulation duration would likely not yield significant differences. The results obtained with AL-MTP scheme were compared with simulations performed using the modified embedded atom method (MEAM) potential [226, 227] (for the radial distribution function and volume) and the property-labelled materials fragments (PLMF) [245] framework implemented in AFLOW-ML [246] (for the mechanical properties), using the same exact conditions. The structural details, including the atomic positions and lattice vectors of the predicted structures, are passed directly as input for the AFLOW-ML module. These calculations are referred to as 'AFLOW-ML' and 'MEAM' in the following.

5.2 Results and Discussion

In order to predict potentially novel materials and assess their stability, the set of candidates found through the AL-MTP pipeline is relaxed on-the-fly, post-relaxed with

DFT, and then the phonon density of states and the respective phonon dispersion are calculated. Specifically, the following steps are followed:

1. calculation of the convex hulls for all binaries and generation of the MTPs,
2. identification of potentially novel stable binaries located at the edges of the respective convex hull,
3. assessment of stability through phonon modes analysis,
4. use of MTPs in MD simulations and calculation of mechanical properties.

The large number of candidates summarized in Tab 5.1 and 5.2 for the complexes Cu-Si, Ni-Si, Cr-Si, Cr-Ni, Cu-Ni, Cu-Cr and Cu-Ni-Si-Cr were used as input in step 1 above. For each of the potentially stable candidates, 12 or fewer atoms were considered in the unit cell for the binary case and 8 or less atoms for the quaternary case. The whole concentration range of the species concentration in the binaries was then scanned and the respective convex hulls were generated. For each of the systems listed in Tab. 5.1, binary MTPs for both the fcc and bcc symmetries, as well as combined, and quaternary MTPs, were generated (step 1 above) with $\mathrm{lev}_{\mathrm{max}}$: 16, 18, and 20 to assess the accuracy of the algorithm and select the most suitable potential. Although $\mathrm{lev}_{\mathrm{max}} = 20$ provided more accurate results for fcc and bcc, $\mathrm{lev}_{\mathrm{max}} = 16$ with around 200 Θ parameters was able to find the same convex hull vertices while being computationally less expensive. For the case of fcc and bcc combined, even the MTPs with $\mathrm{lev}_{\mathrm{max}} = 20$ made drastic extrapolations around the A-rich region of the binary system. Accordingly, with the current framework, two sets of MTPs are constructed for the fcc and bcc structures, respectively. Specifically, at step 1, two sets of MTP potentials were trained for each system, splitting the training set into fcc and bcc ones. This way, it is possible not only to use the convex hull for analysing the relative stability of all relaxed structures but also to assess these with respect to the different symmetries.

Due to the energy resolution of meV and the sensitivity of Hess' law to the DFT energies, precision criteria were set within a range of mean absolute error (MAE) in the range of 1-10 meV. Therefore, structures with a formation enthalpy value below the AFLOW convex hull are considered more relevant. The number of training steps for obtaining each MTP is 500. Through this, the prediction of experimentally observed phases and the discovery of new ones were achieved for all the binary systems considered except Cr-Si and Cu-Cr. For the latter, only the already known Cr-Si phases are found. In Tab. 5.4, information on the generated AL-MTP potentials is reported, including the training set size and the MAE for each system, based on the fcc and bcc sets.

System	fcc		bcc	
	train set size	MAE	train set size	MAE
Cu-Si	688	6.046	706	7.890
Ni-Si	643	8.678	687	11.349
Cr-Si	611	14.022	720	21.198
Cr-Ni	648	5.088	957	10.996
Cu-Ni	313	0.884	1090	2.862
Cu-Cr	301	9.782	457	13.524

Table 5.4 Fitting (MAE in meV/Atom) errors during the AL-MTP process and the generation of the MTPs for the fcc and bcc sets, respectively. The number of configurations selected for the training set are given.

lev_{max}	train set size	MAE	RMSE (σ)
16	1858	9.265	12.178
18	2238	8.617	11.784
20	2731	7.782	10.471

Table 5.5 Fitting (MAE and RMSE in meV/Atom) errors during the AL-MTP process and the generation of the MTPs for the fcc-derivative sets for different lev_{max}: 16, 18 and 20, respectively. The number of configurations selected for the training set are given.

Specifically, the size refers to the number of pre-selected data in the fcc and bcc configurations, as discussed in step (2) of the AL-MTP approach above. As observed in Tab 5.4, the number of pre-selected configurations included in the training set during the relaxation process shows the potential of the AL algorithm. For example, in the Cu-Si system the total number of training configurations is 1,394. Accordingly, the AL-MTP approach enabled a relaxation of all the structures by only running 1394 static DFT calculations instead of 23,712 relaxations as in the size of the total data set in Tab. 5.1. As expected, the fitting MAEs for the systems Ni-Si and Cr-Si are higher than for the other systems, as the respective energy range of the stable structures is of the order of 300-400 meV. Finally, for the quaternary case, Tab. 5.5 presents collected extrapolation error data. An MTP with $lev_{max} = 18$ was chosen for its balanced accuracy and stability at the MD level.

5.2.1 Prediction of Novel Binary Structures

For the binaries considered in this work (see Tab. 5.1), the convex hulls were calculated, and the relevant structures were post-relaxed with DFT. The respective results are

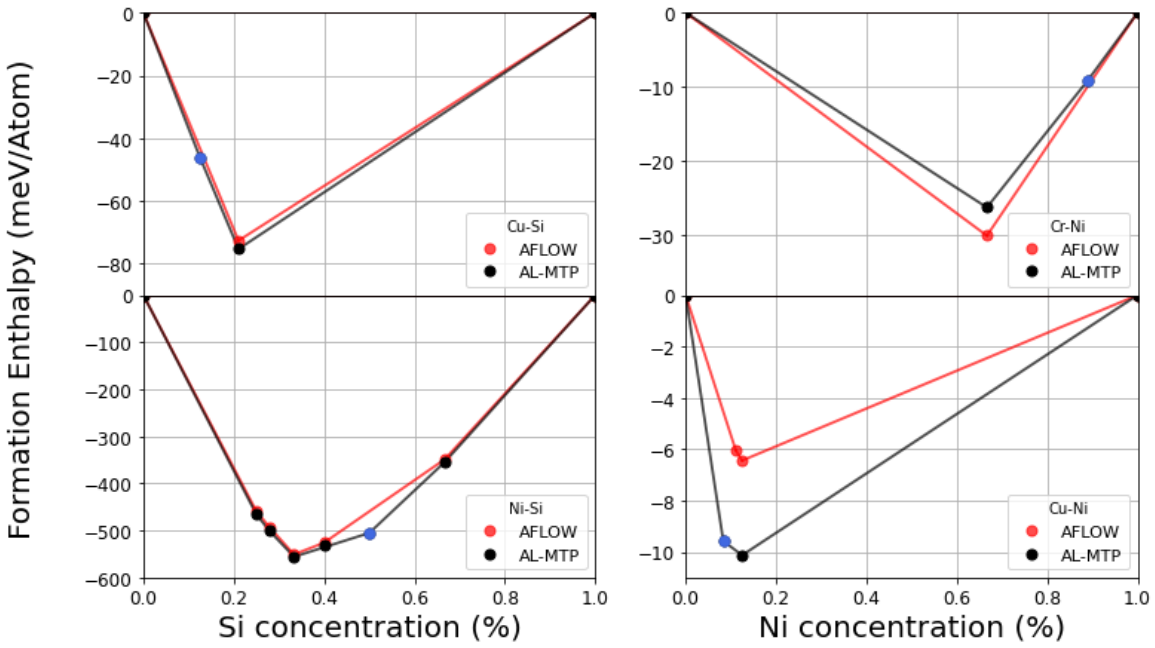


Fig. 5.1 Convex hulls for the binary systems in Tab. 5.1 as calculated in this work are labelled as 'AL-MTP'. The potentially new prototypes are denoted through the blue circles. The respective convex hulls from AFLOW are also provided for comparison [11].

depicted in Fig. 5.1. For comparison, the convex hulls reported in the AFLOW library are also provided. The novel binary candidates found in this work, indicated by the blue circles, are not included in the AFLOW. Accordingly, comparing the AL-MTP work with the AFLOW data, it can be concluded that four new potentially stable binary candidates have been found. These are summarized in Tab. 5.6. The stability of the novel binary structures predicted in this work was confirmed through phonon modes calculations. One of the respective phonon dispersion diagrams will be presented for one of the predicted post-relaxed binaries in the following. In the Cu-Si system, the phase $Cu_{15}Si_4$ and a novel structure Cu_7Si with 8 atoms in the unit cell were found. The rich region of this phase diagram is very complex, with more than four different phases coexisting at low silicon concentrations. The stability of this novel binary phase was confirmed through the phonon dispersion calculated for the post-relaxed structure in Fig. 5.2. Note that the phonon dispersion was also calculated including spin-polarization for the Ni-Si system, confirming that no imaginary frequencies are observed. No imaginary frequencies were found for this novel prototype. In particular, the phonon frequencies obtained using the MTP potentials are in good agreement with the DFT results, providing additional evidence of the accuracy of the trained MTPs in

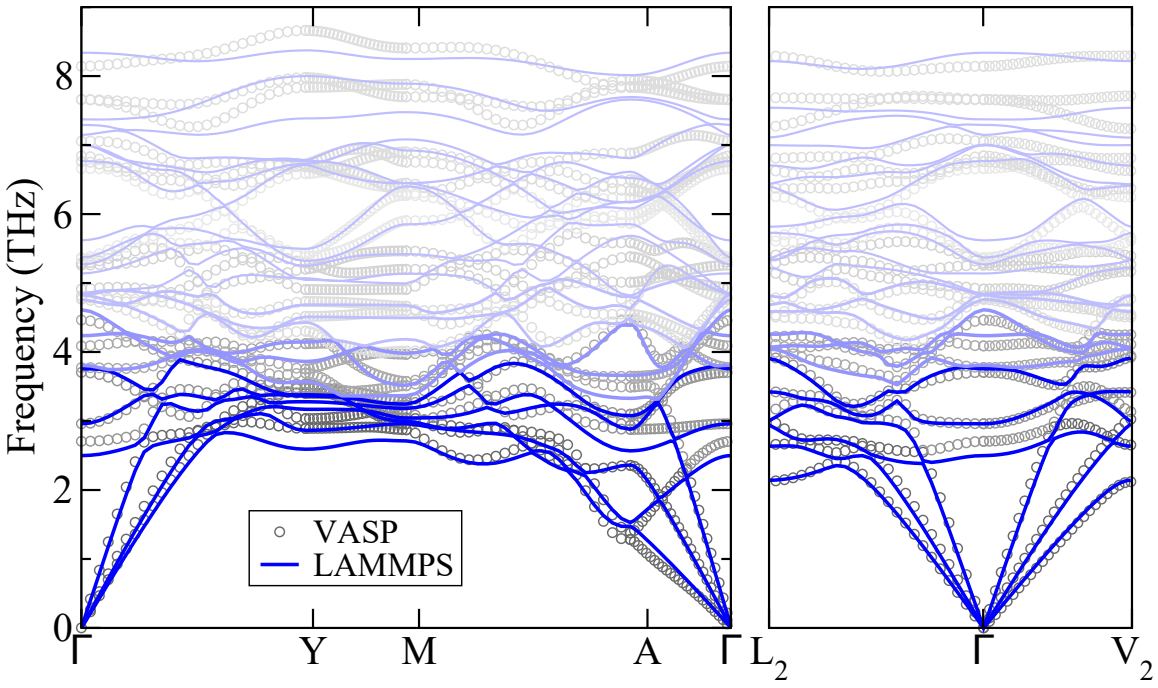


Fig. 5.2 The phonon dispersion for the predicted novel Cu_7Si binary. The phonon spectra calculated within the DFT framework is compared to those obtained using the AL-MTP potential in LAMMPS-MLIP [11].

this work. On top of the predicted Cu_7Si, the larger unit cells of $Cu_{19}Si_6$, $Cu_{56}Si_{11}$, and $Cu_{33}Si_7$ reported in the literature could not be confirmed, as they were not included in the training set. At a 25% Si concentration, a Cu_3Si prototype was found with a lower formation enthalpy than the corresponding one in AFLOW. This was also the case at a 20% Si concentration in Cu_5Si. However, both of these were found to be unstable. The main reason is that the structure of these prototypes should be more complex [247]. Such a structure, composed of three unit cells with two vacancies of copper for Cu_3Si, ending in the stoichiometry $Cu_{3.17}Si$, was previously proposed [247]. No structural DFT data were found for these binaries.

Phase	Formation enthalpy	Source	set symmetry
Cu_7Si	-46.0	ENUMLIB	bcc
NiSi	-505.4	OQMD	—
$CrNi_8$	-9.2	ENUMLIB	fcc
$Cu_{11}Ni$	-9.6	ENUMLIB	bcc

Table 5.6 Formation enthalpy (meV/Atom) and source of the new prototypes found through the AL-MTP procedure and post-relaxed with DFT.

The Ni-Si system was successfully reproduced by the AL-MTP, with a new stable phase added compared to the ones reported in the OQMD database. This material is composed of eight atoms in the unit cell and is not present in either the AFLOW or the ENUMLIB datasets. Predictions match the OQMD convex hull as well as that of the Materials Project, where the phase Ni_3Si_2 has an energy of 0.006 meV above the hull edges [248]. Accordingly, although this has been experimentally observed, it is thermodynamically metastable or unstable. This is the only black circle that is not part of a convex hull in Fig. 5.1. As an additional check of the prediction ability of the method, confirmation was obtained through MD simulations and the AL-MTP potential that the known stable $Ni_{31}Si_{12}$ structure remains a stable solid structure for temperatures lower than $\approx 1400\,\mathrm{K}$ and has a liquid phase for larger temperatures. The calculated melting temperature is about 100 K larger than the experimental value for the Ni-Si alloy with 28 % Si. The mechanical properties for $Ni_{31}Si_{12}$ were also calculated and compare well with DFT results [249]. These include c_{11}=311.4 GPa (301 GPa), c_{12}=169.6 GPa (163 GPa), c_{44}=68.1 GPa (60 GPa), B = 202.7 GPa (199 GPa), and μ = 68.9 GPa (68 GPa) based on AL-MTP (DFT). The study of the Cr-Si binary using the AL-MTP and DFT post-relaxation did not provide any novel phases but confirmed the convex hull reported in OQMD and AFLOW. Nevertheless, a higher formation enthalpy for $CrSi_2$ was reported [53] using the DFT implementation in CASTEP [250]. This increases the number of vertices of the convex hull from two (Cr_3Si and $CrSi_2$) to four, where Cr_5Si_3 and CrSi now would become stable. Predictions align with the directions indicated by the three repositories previously cited.

Regarding the Cr-Ni system, the experimentally observed phase $CrNi_2$ was predicted by the AL-MTP, as well as the novel candidate $CrNi_8$. $CrNi_3$ was found to have a formation enthalpy of a few meV over the convex hull and was slightly unstable, as observed in AFLOW. The energy value of the well-known $CrNi_2$ of around -30 meV differed substantially when comparing AFLOW with Materials Project's and this work. This phase was not reported in OQMD, where only $CrNi_3$ was found to be stable. Finally, the potentially novel phase $CrNi_8$ has the same structure as a respective unstable phase in AFLOW. The difference in energy in the main stable phase $CrNi_2$ renders $CrNi_8$ also stable. In ENUMLIB, the same structure as in AFLOW for the prototype $CrNi_8$ was found. For the systems Cu-Ni and Cu-Cr, it was not expected to discover any stable phases because they are reported to be simple isomorphous and eutectic systems. Only cluster expansion methods have suggested possible stable phases in the Cu-rich region of Cu-Ni. In this work, Cu_7Ni and a novel candidate $Cu_{11}Ni$ were predicted by the AL-MTP, turning Cu_8Ni unstable compared to AFLOW.

Phase	AFLOW	DFT	AL-MTP
$Cu_{15}Si_4$	-72.6	-75.3	-77.8
Ni_3Si	-459.6	-466.8	-460.4
$Ni_{31}Si_{12}$	-492.6	-501.0	-501.5
Ni_2Si	-550.7	-556.8	-547.1
Ni_3Si_2	-525.0	-530.4	-525.5
$NiSi_2$	-349.8	-350.6	-356.6
Cr_3Si	-353.5	-353.6	-361.4
$CrSi_2$	-376.7	-366.5	-372.2
$CrNi_2$	-30.0	-26.0	-27.2
Cu_7Ni	-6.4	-10.1	-9.1
Cu_8Ni	-6.0	-9.2	-8.8

Table 5.7 Formation enthalpies (in meV/Atom) of stable binary structures included in the AFLOW as compared to the DFT post-relaxed and AL-MTP values.

For the Cu-Cr system, all the 23,708 prototypes had positive formation enthalpy values. This agrees with the experimentally found precipitates in Cu-Ni-Si-Cr alloys, where only Cr-rich clusters are observed and not binary precipitates.

Regarding the efficiency of the AL-MTP method, the number of structures is much higher than those reported in AFLOW. This strongly underlines the possibility of much broader screening of the configurational space in multi-component systems. Some of the structures added to the training set from AFLOW and OQMD were automatically generated also by ENUMLIB. This are for example the Ni_3Si, Cu_7Ni and $CrNi_8$ structures. This repetition was eliminated during post-relaxation. In order to further test the efficiency and potential of AL-MTP, the formation energies predicted by AL-MTP were compared using the DFT values from AFLOW and VASP. Due to the higher number of initial candidates considered, new binary phases for the four A-B systems, mostly in their A-rich region, could be discovered. In Tab 5.7, the formation enthalpies for all the reported stable phases are shown, calculated via AL-MTP for one AL cycle and compared to AFLOW and post-relaxation results in VASP.

Summarizing the AL-MTP predictions: The new prototype Cu_7Si is proposed as a potentially interesting candidate, as there are structures with the same stoichiometry experimentally observed in the literature. The respective unit cell before and after DFT relaxation is depicted in Fig. 5.3. The relaxed geometry of Cu_7Si is considerably different from the prototype generated by ENUMLIB and any of those in the initial pool of candidates. This is due to the AL-MTP model flexibility, compared to on-lattice methods such as cluster expansion. According to AFLOW-SYM [251], the crystal structure Cu_7Si found by AL-MTP is orthorhombic-bipyramidal and has a space group

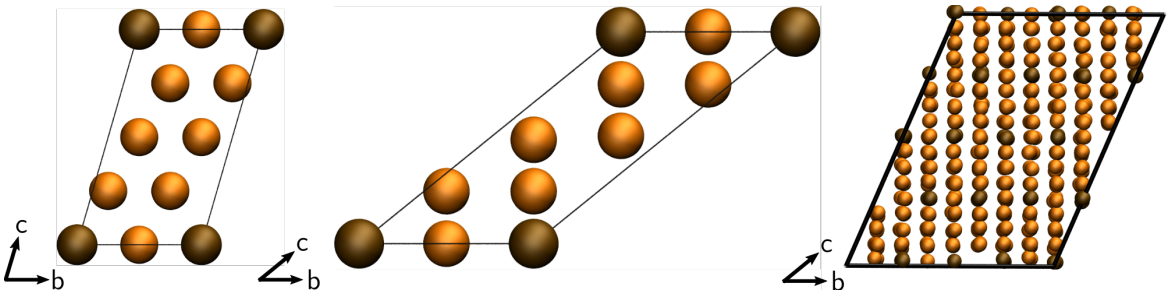


Fig. 5.3 Crystal structure of the novel predicted Cu_7Si before (left) and after (middle) relaxation. The Cu and Si atoms are colored in orange and brown, respectively. The right panel depicts the Cu_7Si supercell at a temperature $T = 300\,\mathrm{K}$ [11].

Phase	a	b/a	c/a	α	β	γ	Vol	LS	SG
Cu_7Si	4.460	1.000	2.075	61.5	119	70.2	96.35	ORCF1	*Fmmm*
NiSi	5.140	0.654	1.090	90	90	90	96.83	ORC	*Pnma*
$CrNi_8$	3.506	1.582	1.582	95.7	108	71.6	97.08	BCT1	*I4/mmm*
$Cu_{11}Ni$	5.128	0.706	1.732	114.1	106.8	90	142.98	ORCC	*Cmmm*

Table 5.8 Lattice vectors (Å), angles (°), volume (Å^3), lattice system (LS) and space-group (SG) for the new structures.

69-*Fmmm*. This differs from the high-temperature κ-Cu_7Si CsCl-type phase previously proposed [252] and from the Cu_7Si prototype with the lowest formation enthalpy found in AFLOW. The latter is cubic with an fcc Bravais lattice. At 0 K, the AL-MTP structure is lower in enthalpy by -8.1 meV/atom than the one from AFLOW. As these two stable Cu_7Si structures differ in symmetry, they are also expected to show distinct mechanical and thermal properties. However, — to my knowledge — , such data are not available in the literature to provide a comparison with AL-MTP computations. The NiSi phase was previously proposed in OQMD and by experiments. For $CrNi_8$, the energy below the hull is too small for this to be considered a technologically relevant structure. In addition, no structure with this stoichiometry has been experimentally observed. This is also the case for the predicted new $Cu_{11}Ni$, where only cluster expansion methods and chemical similarity methods have predicted Cu-rich stable phases in Cu-Ni. In Tab. 5.8, the main structural parameters of the most relevant new prototypes are summarized.

Spin-polarization Effects

In order to test the applicability of the AL-MTP algorithm on binary alloys with magnetic alloying elements, the contribution of magnetic moments is assessed. Spin

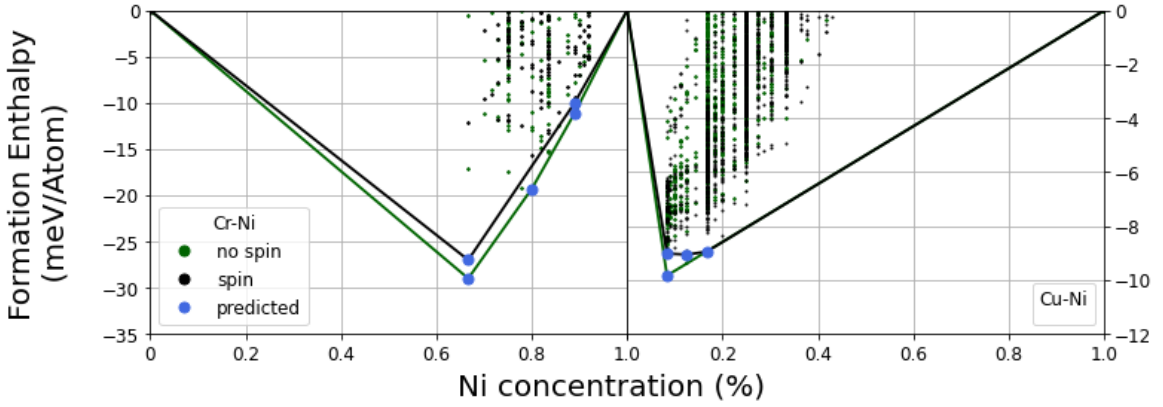


Fig. 5.4 The convex hulls for the Cr-Ni (left) and the Cu-Ni (right) binaries. The black and green lines correspond to the spin-polarized and spin non-polarized calculations. The blue circles denote the predicted structures [11].

effects can cause notable shifts in energy calculations, particularly with AL-MTP's extrapolations. In such a case, the vertices of the convex hull would not be properly predicted affecting the prediction of novel structures. In order to assess whether significant energetic shifts are induced, the convex hull calculation for the binaries Ni-Si, Cr-Ni, and Cu-Ni was repeated, including spin polarization in the DFT simulations run by AL-MTP. The formation energies in the bulk fcc crystal of Ni and bcc crystal of Cr were first compared via DFT relaxation. The formation energies for the fcc Ni crystal were calculated at -5.5670 eV/atom and -5.5242 eV/atom with and without spin polarization. In bcc Cr, these are -9.6318 eV/atom and -9.6318 eV/atom, respectively. The results match previous findings on the difference between the formation energy of fcc-Ni with and without spin polarization in the range of 50-60 meV/atom [52]. No relevant shift in the energy value for bcc-Cr was found. The convex hulls for both spin-polarized and non-polarized calculations presented in this section use the fcc-Ni energy value predicted by AL-MTP as a reference. For the Ni-Si binaries, no significant influence was observed. As expected, the predicted energy from the spin-polarized convex hull is lower for the structures with local magnetic moments around the Ni-rich region.

In Fig. 5.4, the convex hulls for the Cr-Ni and Cu-Ni binaries, with and without spin polarization, are depicted. In these binaries, an increase in the formation energies of around 1-5 meV/atom was observed, which shifted the edges and corners of the respective convex hulls to higher energies of that order compared to the spin non-polarized case. As observed, in Cr-Ni, the convex hull is slightly deeper in energy without including spin polarization, with a shift in the formation enthalpy for $CrNi_2$.

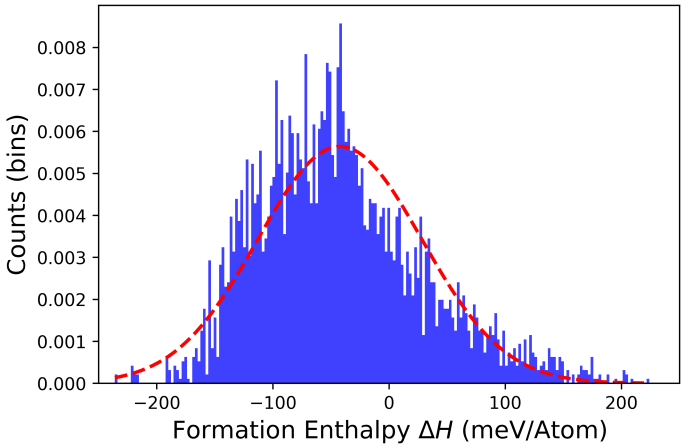


Fig. 5.5 The number of predicted quaternary structures with respect to their formation enthalpy ΔH (meV/Atom).

This effect is assigned to the fact that the energy difference between spin-polarized and spin non-polarized is lower than the AL-MTP MAE from Tab. 5.4. This energy difference can lead to a change in the number of convex hull vertices, modifying the potentially novel stable structures to be post-relaxed. Specifically, the energy shift affects the stability of the predicted $CrNi_4$, which becomes unstable when adding spin polarization. However, the formation enthalpy of the newly predicted $CrNi_8$ remains below the convex hull edge when adding spin polarization. Due to the lower formation enthalpy calculated for the new $Cu_{11}Ni$, the stability of Cu_7Ni reported in AFLOW and ENUMLIB was not confirmed by the spin non-polarized calculations for the Cu-Ni binaries. Again, the enthalpy difference for this structure is lower than the MAE for Cu-Ni. The spin polarization was recently included in a new version for generating magnetic MTPs (mMTP) [253], taking into account the magnetic degrees of freedom. However, this version was and still is not available for extending the current work. Accordingly, it is concluded that, though MTP is not capable of mapping the energy landscape for structures with local magnetic moments with high accuracy as compared to mMTP, there is no critical difference between spin-polarized and spin non-polarized cases for the AL-MTP convex hulls calculated here, allowing further exploration in view of discovering new structures.

5.2.2 Analysis of Metastable Quaternary Structures

At a temperature of 500 °C, a thorough examination of a Cu-Ni-Si-Cr alloy was conducted through high-resolution transmission electron microscopy and scanning

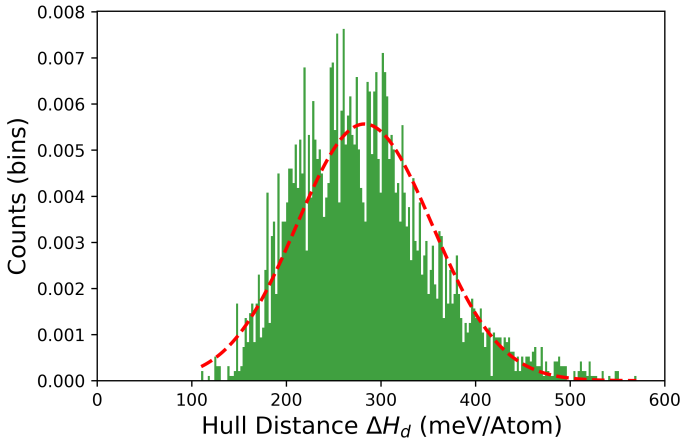


Fig. 5.6 The number of predicted quaternary structures with respect to their hull distance ΔH_d (meV/Atom).

transmission electron microscopy, unveiling the simultaneous existence of three distinct precipitates [38]. These identified phases manifest as the ordered face-centered cubic β-Ni_3Si, the orthorhombic δ-Ni_2Si, and an ordered face-centered cubic (Ni, Cr, Si)-rich phase. This section explores the latter quaternary phase and its metastability. In order to screen the energy landscape of the generated quaternary dataset, the generated structures from Tab. 5.2 are first relaxed via the AL-MTP pipeline, and their formation enthalpies are shown in Fig. 5.5.

One of the most important thermodynamic quantities that controls phase stability is the distance hull ΔH_d or decomposition energy [254]. The energy distance of a compound to the convex hull is hence a measure of its instability. ΔH_d serves as a measure of the stability of a material providing valuable insights into both the assessment of its stability and the feasibility of synthesis. Additionally, it offers useful information concerning the uncertainties associated with these factors, further enhancing our understanding of the material's overall characteristics. The distance to the convex hull ΔH_d is a measure of thermodynamic stability. The probability to synthesize a material decreases rapidly with ΔH_d, which gives an idea or how metastable a structure is. In Fig 5.6, the histogram for the hull distances computed are depicted. Given the results, stable quaternary structures are not anticipated. Instead, we can expect metastable candidates within the 100-150 meV/Atom range. These structures are post-relaxed to confirm their thermodynamically stability and to identify the best candidate by comparing them with the experimentally observed fcc structure.

In the end, the post-relaxation process yielded a number of potentially metastable candidates, with the most relevant ones emerging from the stoichiometries $Cu_2Ni_2Si_2Cr_2$

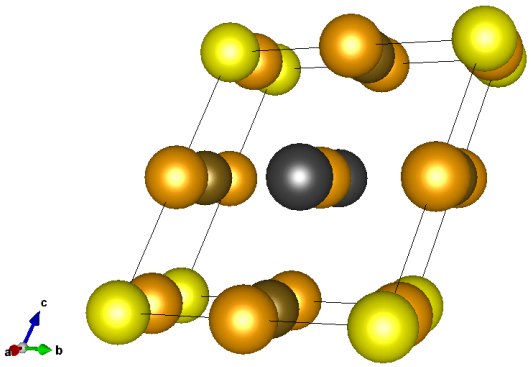


Fig. 5.7 Crystal structure of the novel predicted Cu_4NiSi_2Cr. The Cu, Ni, Si and Cr atoms are coloured in orange, brown, yellow and dark grey respectively.

Phase	AL-MTP	DFT
Cu_4NiSi_2Cr	-75.8	-77.5
$Cu_2Ni_2Si_2Cr_2$	-244.5	-247.6

Table 5.9 Formation enthalpies (in meV/Atom) of metastable quaternary structures. Comparison between the AL-MTP and DFT post-relaxed values.

and Cu_4NiSi_2Cr. These correspond to the lowest formation enthalpy among all the stoichiometries from Tab. 5.2 ($Cu_2Ni_2Si_2Cr_2$) and to the prototype with the lowest formation enthalpy within Cu_4NiSi_2Cr. The formation enthalpies for both prototypes are presented in Tab. 5.9. The stoichiometry Cu_4NiSi_2Cr closely aligns with the experimental observations of the fcc (Ni, Cr, Si)-rich phase, as documented in [38]. The STEM/EDS microanalyses of the small precipitate revealed a composition of 50.52% Cu, 11.84% Ni, 15.94% Si and 9.04% Cr, corresponding to Cu_4NiSi_2Cr stoichiometry. In Fig. 5.7, the novel Cu_4NiSi_2Cr structure is depicted, with its structural details outlined in Tab 5.10.

Phase	a	b/a	c/a	α	β	γ	Vol	LS	SG
Cu_4NiSi_2Cr	5.219	1.000	1.000	95.63	123.31	56.69	94.97	BCT	*I4/mmm*

Table 5.10 Lattice vectors (Å), angles (°), volume (Å^3), lattice system (LS) and space-group (SG) for the novel Cu_4NiSi_2Cr.

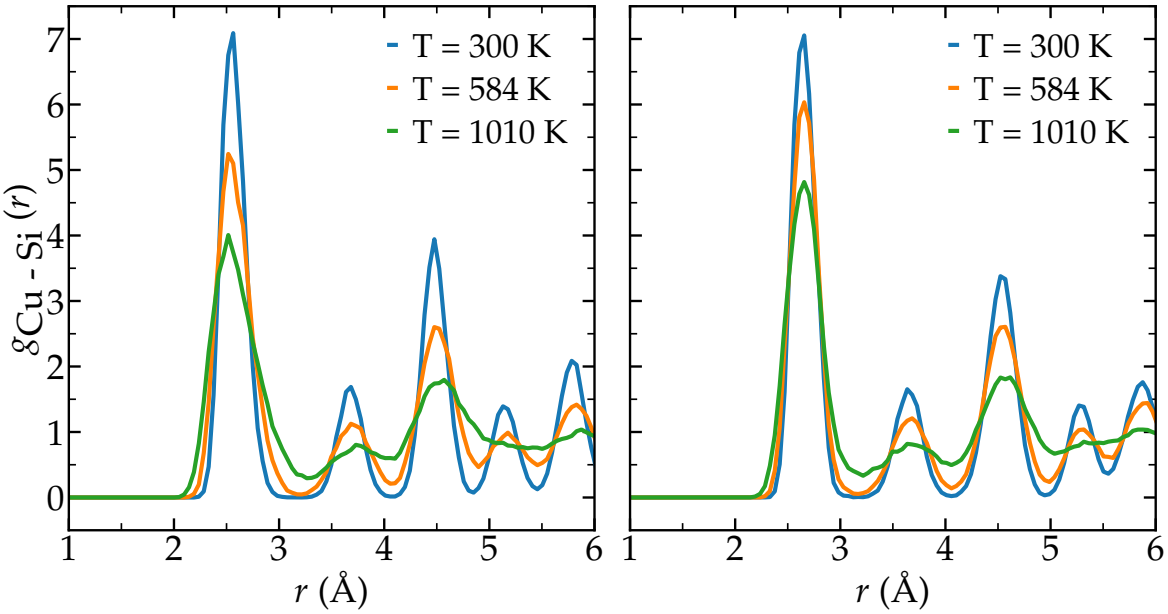


Fig. 5.8 The RDF for the Cu-Si pair calculated through MD simulations using the AL-MTP (left) and MEAM (right) potentials, respectively. The curves correspond to simulations at different temperatures T, as denoted by the legends [11].

5.2.3 Properties Assessment of the Predicted Cu_7Si

In order to first assess the efficiency and accuracy of the MTPs and, most importantly, reveal their practical importance, these MTPs are used for the calculation of material properties. Specifically, the MTPs generated in this work are further implemented in LAMMPS to perform MD simulations. This section focuses on the novel Cu_7Si binary structure predicted above, and its structural, mechanical, and thermal properties are calculated. The supercell of Cu_7Si used for the MD calculations is depicted in Fig. 5.3. As a measure of the structural properties and the influence of thermalization, the RDF at different temperatures is computed, as summarized in Fig. 5.8. In this figure, the results from both the AL-MTP and the MEAM for the Cu-Si RDF are shown. At 300K, well defined peaks at the crystal position of the nearest-neighbors can be observed. For both potentials, the RDF trends indicate a melting of the Cu_7Si structure at a temperature in the range 584 K and 1010 K. Additionally, the well-resolved RDF suggests that the signal-to-noise ratio is excellent, further supporting the adequacy of the 200 ps production run. This is consistent with the Cu-Si phase diagram, which indicates a melting temperature for Cu_7Si in the range $800 - 1400$ K [255]. The respected changes in the density of the binary with the temperature are depicted in Fig. 5.9. The density of Cu_7Si decreases with the temperature revealing

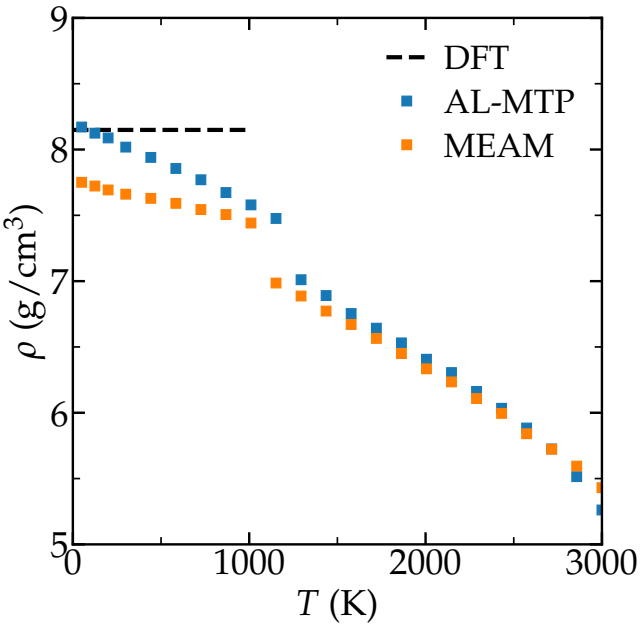


Fig. 5.9 Density of the Cu_7Si structure as a function of the temperature T, for both the AL-MTP and MEAM potentials as indicated by the legend. The horizontal dashed line highlights the value $\rho = 8.15\,\mathrm{g/cm^3}$ as measured from DFT in the limit $T \to 0\,\mathrm{K}$ [11].

an extension of the volume of the box of 6% in the range [300-1000] K and 15% in the range [1000-2000] K as calculated with AL-MTP. The respective MEAM values are 3% in the range [300-1000] K and 16% in the range [1000-2000] K. An abrupt decrease in the density is observed for both potentials (at a temperature of 1200 K and 1000 K for AL-MTP and MEAM, respectively). This abrupt transition is indicative of a considerable structural change in the crystalline structure of the Cu_7Si binary, which is consistent with the melting behavior observed in the RDFs of Fig. 5.8, as well with the expected melting temperature of Cu-Si systems [255]. Overall, both AL-MTP and MEAM potential give similar qualitative behavior, and both predict a melting near 1000-1200 K, which confirm that the structures and phases predicted using AL-MTP are reasonable. Moreover, the better agreement between the AL-MTP potential and DFT in the limit $T \to 0$ suggests that our potential captures the low temperature limit better than the MEAM force field. Regarding the abrupt transition at around $T = 1200K$, this is related with a crystal-to-amorphous phase transition in the material.

In order to assess the elastic response of the predicted stable Cu_7Si binary, elastic constants, as well as the bulk and shear moduli, are calculated. The respective simulations are again performed twice, using the AL-MTP and the MEAM potentials.

According to the AL-MTP (MEAM) calculations the main elastic constants for Cu_7Si binary are c_{11}=194.5 GPa (222.5 GPa), c_{12}=81.7 GPa (109.2 GPa), and c_{44}=35.7 GPa (33.3 GPa). The calculated values for the bulk and shear moduli are $B = 124.1$ GPa (153.6 GPa) and $\mu = 41.8$ GPa (39.9 GPa), respectively. For all moduli the numbers in parentheses correspond to the values calculated with the MEAM potentials. The trends found among these moduli are confirmed by both the AL-MTP and MEAM calculations. Note, though that previous studies comparing MEAM to DFT calculations for Cu-Si binary compounds such as Cu_2Si or Cu_3Si [256] suggest that MEAM potentials are not the most suitable for the calculation of Cu-Si mechanical properties. Accordingly, the moduli calculated with AL-MTP are expected to be more accurate than the MEAM ones. This is further supported by the very good comparison of the AL-MTP calculated moduli with those from AFLOW. The latter ones are $B = 122.5$ GPa and $\mu = 41.4$ GPa. Accordingly, the AL-MTP potentials of this work are of high practical importance for two reasons: the computational efficiency in their generation and the related accuracy they add in simulations as compared to other interatomic potentials.

5.3 Summary

In this work, the efficiency and reliability of MLIPs for identifying new materials and phases for copper alloys have been discussed. Specifically, with Cu alloys of the form Cu-Ni-Si-Cr in mind, MTPs are derived from QM simulations using AL. For the alloy-relevant binary structures, the formation enthalpies of thousands of structures have been calculated, and the convex hulls constructed. This allowed the discovery of new stable binary structures not reported previously. Simultaneously, previously reported stable phases have been predicted, highlighting the applicability and efficiency of the method. The stability analysis of the predicted new binaries was extended by checking the phonon dispersion for imaginary frequencies. The stable nickel silicide NiSi was identified, already reported in OQMD but not in AFLOW. The Cr-Ni convex hull suggested a potential stoichiometric compound on the line of the predicted candidate $CrNi_8$. However, the study of intermetallic phases in the binary complexes Cu-Ni and Cu-Cr is not relevant beyond academic interest. For the Cu-Ni binaries, a new stable phase $Cu_{11}Ni$ was found. For the Cu-Cr and Cu-Ni binaries, no stable structure could be observed, in accordance with experimental observations. From all predicted structures, as practically more relevant, the focus was further directed to the predicted stable Cu_7Si binary. For this, the generated AL-MTPs were implemented in MD simulations to calculate its structural, thermal, and elastic properties. These

were found to be more accurate than those calculated with other classical interatomic potentials and showed similar accuracy to other ML-derived results. For the quaternary system Cu-Ni-Si-Cr, the metastable structure Cu_4NiSi_2Cr generated from an fcc parent lattice was predicted, aligning with the experimentally observed fcc (Ni, Cr, Si)-rich phase.

In the end, a flexible AL approach was applied, assessed, and used efficiently to implement QM accuracy for predicting new materials and calculating their properties. Although the focus was on a certain class of materials related to Cu alloys, the findings underline the applicability of AL-MTPs in materials discovery. Importantly, this approach efficiently modelled configurations initially not included in the pool of training structures. In addition, the AL cycle could train the potentials automatically and construct convex hulls with just a small number of single-point DFT calculations compared to the number of candidates that need to be relaxed when using other methods. In this work, an MTP of $lev_{max} = 16$, trained solely with 7,821 single-point DFT calculations, successfully relaxed 142,267 prototypes for the binary case. In the quaternary system, only 2,238 single-point DFT calculations were needed to train an MTP of $lev_{max} = 18$ for the relaxation of 4,159 candidates. Indeed, this method is applicable to more complex systems, such as quaternary alloys, currently being investigated within the framework of a follow-up study. In this way, the generation of both the potentials and the stability curves can be significantly accelerated. The further use of the generated AL-MTP, together with MD simulations, also enables a significant increase in spatiotemporal scales in the simulations, allowing the modeling of much larger supercells for longer times. In this way, the simulations gradually move closer to experimental scales. In the end, the results provide an efficient and systematic recipe for detecting new stable precipitates in alloys and generating accurate, transferable MLIPs that can be further used in MD simulations for larger-scale investigations of the mechanical properties of precipitates in multi-component alloys.

5.4 Acknowledgement

Most part of this chapter was reproduced from Ref [11] with permissions from Materials Chemistry and Physics, Elsevier Journal.

Chapter 6

Enhancing Nanopore Translocation Read-out via PDML

In this chapter, a novel ML framework for analyzing ionic and transverse currents in nanopore experiments, facilitating their simultaneous examination in nanopore experiments is introduced. This breakthrough opens the door to more intuitive base-calling methods, enabling the efficient identification of molecules as they pass through nanopores via electrophoresis. The approach has been rigorously tested and evaluated, with a focus on biomolecule translocation through 2D nanopores, marking a significant advancement in nanopore technology.

6.1 Data Collection

6.1.1 Experimental Details

In this study, current traces from previous nanopore experiments [257, 258] were utilized. The data, obtained from two main sources — DNA nucleotide and different bio/molecule translocation experiments through two-dimensional molybdenum disulfide (MoS_2) nanopores — was collected and provided by the Radenovic group at the Laboratory of Nanoscale Biology, EPFL, Switzerland. Simultaneous measurement of ionic and electronic signals allows the detection of various analytes based on the current blockades they induce. These blockades correspond to the translocation events, which are further analyzed in this work and carry the identity of the translocating bio/molecule. For the single DNA nucleotides dAMP, dTMP, dGMP, and dCMP, two sets of single channel (ionic) experiments were conducted, each with different nanopore diameters: experiment A (Exp. A) with a nanopore size of 3.3 nm and experiment B

analyte	pore [nm]	*cis-trans* [M]	salt	V_{ionic} [V]	V_{el} [V]
Nucleotides (d[X]MP)	3.3	0.1 / 0.1	KCl	-0.2	-
Nucleotides (d[X]MP)	2.8	0.1 / 0.1	KCl	-0.2	-
'ssDNA' (80nt ssDNA)	5.9	0.01/0.1	KCl	1	+0.5
'dsDNA' (1000 nt dsDNA)	2.5	0.01/1	KCl	1	-1.4
'polyLys' (30-70 kDa polylysine)	2.5	0.01/1	KCl	-0.5 V, 0 V	-1.2

Table 6.1 Overview of the most relevant experimental details and conditions related to the analyzed data. 'Analyte', 'pore', 'salt', V_{ionic} and V_{el}refer to the translocating molecule, the pore diameter, the salt solution, the voltage difference in the ionic and electronic channel, respectively, and the presence of a differential amplifier.

(Exp. B) with a nanopore size of 2.8 nm. For the bio/molecules, datasets from three different experiments involving biopolymers electrophoretically driven through the MoS_2 nanopores are analyzed using a double channel (ionic and electronic) approach. The biomolecules include a one thousand base-pairs double-stranded DNA (1 kb dsDNA), a negatively charged 80 nucleotides single-stranded DNA molecule (80 nt ssDNA) at a 0.01M concentration, and a positively charged poly-D-lysine hydrobromide (polylysine) with an average molecular weight (mw) of 30×10^3 to 70×10^3 g/mol. The labels 'ssDNA,' 'dsDNA,' and 'polyLys' are used for these experiments and analytes, respectively. It is noted that the two chambers of the experiments, the *cis* in which the biomolecules are placed before translocation and the *trans*, which is the chamber in which the biomolecules enter after their translocation through the pore, are filled with the same salt at different concentrations to assist the electrophoretic motion of all species (anions/cations and biomolecules) through the pore. Additionally, in the polyLys experiment, two different ionic voltages were applied for the same purpose. All experimental details are provided in the original publication for the nanopore experiments [258, 257]. The conditions of all the experiments such as concentration, viscosity gradient, and voltage difference across the pore are referenced in Tab. 6.1.

6.1.2 Event Detection

To detect translocation events in the raw ionic current data, the cumulative sum (CUSUM) algorithm, implemented in the Open Nanopore software [120], was utilized. This same algorithm was applied to smoothen the raw signals, resulting in piecewise constant currents known as the levels of an event. The CUSUM algorithm fits the raw signal data from the experiments and generates structured data files containing various essential information, such as the concatenated raw signal, the time unit, the sampling frequency, and the event database. Within the event database, details such

Nucleotide	Exp. A	Exp. B	Total	Filtered
dAMP	22	3887	3909	3887
dCMP	391	673	1064	1063
dGMP	757	121	878	789
dTMP	240	127	367	367
ssDNA	-	-	700	585
dsDNA	-	-	100	50
polyLys	-	-	417	188

Table 6.2 Dataset sizes from the two experiments (Exp. A and Exp. B with nanopore diameter 3.3 nm and 2.8 nm, respectively). The left columns refer to the initial nucleotide data, while the right two columns ('Training set A' and 'Training set B') refer to the nucleotide data after the detection of outliers. The ionic current blockades are given in nA, the dwell times in ms.

as the start and end points for each event, the number of levels, the dwell time (time of residence in the pore), and the fitted ionic and electronic current blockade values during the translocation events can be found. In short, this algorithm, details of which can be found elsewhere [259], attempts to detect abrupt changes in the current, allowing an easy identification of the different levels in a translocation event. From each nucleotide experiment [257], four distinct datasets featuring raw signals resulting from the translocation of four single DNA nucleotides: dAMP, dTMP, dGMP, and dCMP, were obtained. In total, there were eight different datasets (see Tab. 6.2). It is worth noting that these datasets vary significantly in terms of the number of samples from each experiment and each nucleotide, posing a substantial challenge during the learning process. From the bio/molecule translocation experiments and the application of CUSUM, 700, 100, and 417 events were detected for the ssDNA, dsDNA, and polyLys experiments, respectively. Note that in this case, the ionic and electronic blockades are correlated in time.

6.2 Feature Extraction

The concatenated translocation events as produced by CUSUM are used for further processing. Representative concatenated signatures are shown in Fig. 6.1. In order to reduce the dimensionality of the input, certain features have been extracted. These features were selected considering physically intuitive aspects, aiming to include information on the dynamics of the translocation process. The ionic current raw signal from the events is encoded into four features. According to the previous work [122],

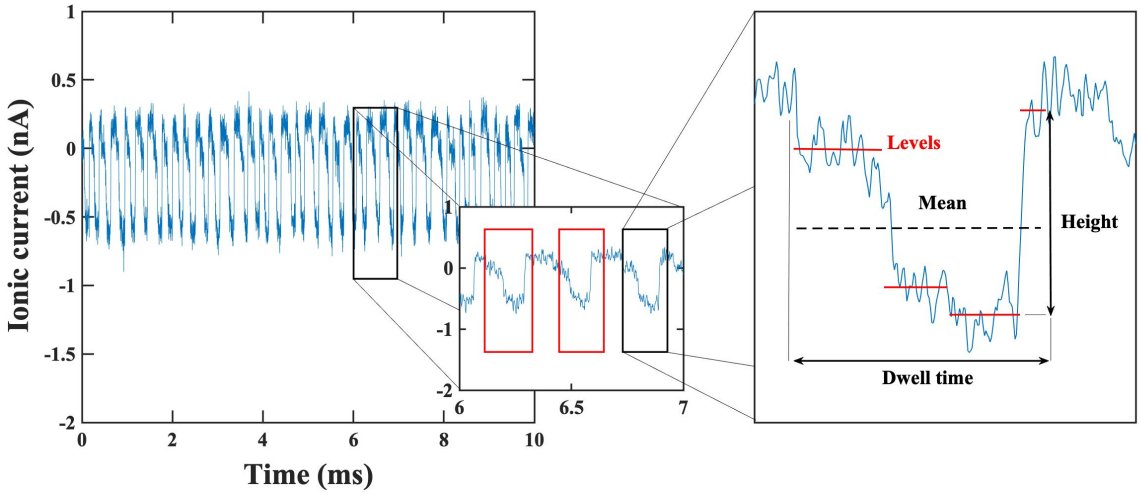


Fig. 6.1 A set of concatenated events for the translocation of dAMP through the nanopore with a diameter of 2.8 nm (Exp. B) . Each red block on the left represents an event of a nucleotide translocating the nanopore in a certain configuration. On the right, the four features for a single nucleotide translocation event are highlighted [12].

the following four features are used here: (a) the dwell time (i.e., the duration of a translocation event), (b) the height of the ionic current blockade (i.e., the difference between the fitted maximum and minimum values of a single current blockade peak), (c) the ionic blockade mean current (mean) (i.e., the average current value), and (d) the levels (i.e., the number of distinct current jumps within a single translocation event). These features are labeled in Fig. 6.1. In parentheses, the notation of the respective ionic/electronic features as used in the following is given. In this notation, e.g., ionic height ('height$_i$') and electronic mean ('mean$_{el}$') correspond to the height in the ionic channel and the mean in the electronic channel, respectively. The features were extracted from all data, i.e., all experiments and both detection channels (ionic and electronic) when applied. The features used in the analysis are those already proven as very efficient data indicators from previous studies [259, 260]. Specifically, the features height and mean could very clearly cluster the translocation events in a two-dimensional feature space, denoting distinct types of translocating molecular conformations [122]. The type of features extracted, based on the definitions above, are visualized on the processed data (after the application of the CUSUM algorithm, red lines for the feature Levels) left side of Fig. 6.1. Regarding the information on the pore diameter, this is implicitly included in the selected features as it leads overall to different values in the extreme values of the ionic/electronic traces. Larger pore sizes provide more noisy signals molecule events due to the smaller region the molecule

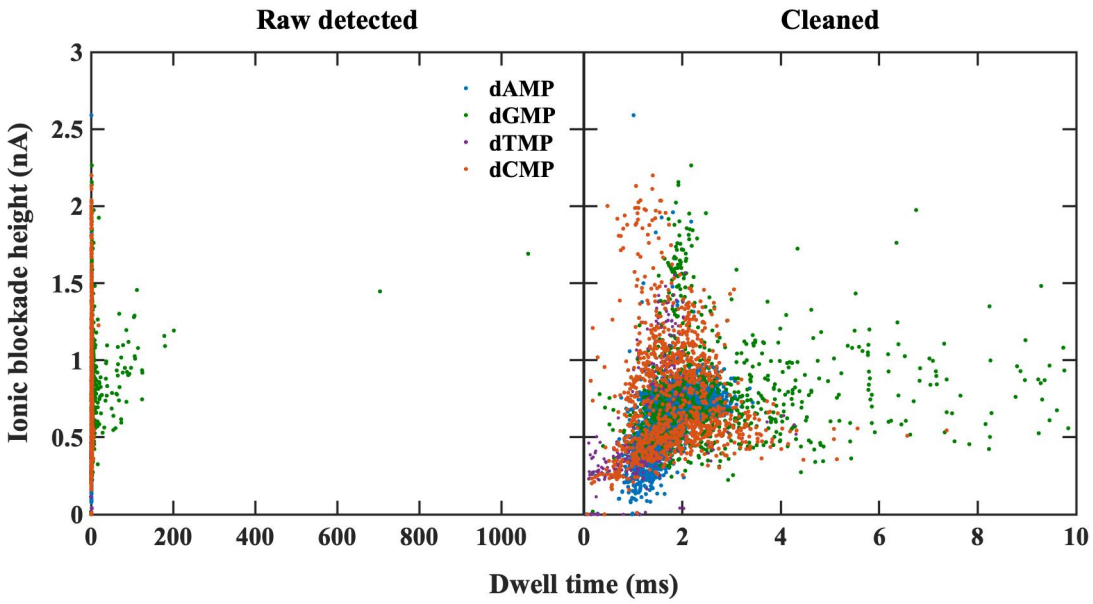


Fig. 6.2 Detection of outlier events in the data from both experiments Exp.A and Exp.B. The left panel depicts all data with respect to the two features of the ionic current blockade and the dwell time, while the right panel reveals the same feature space after the outliers have been remove based on the cutoff for a dwell time below 10 ms [12].

blocks during translocation. Accordingly, in the learning process, the pore diameter is implicitly also learned.

Outlier Detection

In order to reduce the noise and improve the quality of the datasets, two main approaches for the nucleotides and the bio/molecules were followed: dwell time threshold and CUSUM filtering, respectively. For the nucleotides, the concatenated data series revealed outlier events, some of which strongly deviated not only from the ionic current values of the other events but also from events obtained under very similar experimental conditions. Outlier events were also related to a much longer dwell time, probably stemming from stagnation within the nanopores or even retraction events, and are, in any case, not physical, i.e., they do not correspond to translocation events. Possibly, oversensitivity of the CUSUM algorithm could lead to outlier events. In any case, previous analysis under the same experimental conditions has shown that the expected dwell times are within the range 0 – 2 ms [257]. Based on all this, a threshold for excluding outlier events was set at 10 ms. In Fig. 6.2, the result of the nucleotide outlier

dataset	ionic			electronic		
	a	S	E	a	S	E
ssDNA	0.9999	0.5	4.25	0.9999	0.5	4.25
dsDNA	0.9999	0.25	4	0.99	-0.5	2.5
polyLys	0.9999	0.5	2.75	0.99	-0.5	2.75

Table 6.3 Filter parameters for the CUSUM algorithm, applied for the event detection in both ionic and electronic channels.

detection is depicted. In order to minimize noise in the ionic/electronic current traces originating from bio/molecules and ensure clear detection of translocation events, a filter implemented in the software was employed. The application of this filter coefficient "a," which attempts to further process the data for feature extraction, reduced the detected translocation events to 585, 50, and 188 events for the ssDNA, dsDNA, and polyLys experiments, respectively. Higher values of the related filter coefficient (a) will detect fewer events, as expected, while lower values would interpret a reasonable fluctuation in the current as an event. In addition to this filtering, the start (S) and end (E) thresholds defining an event must be set accordingly by matching events in both electronic and ionic channels. Different values for these threshold frames were used in the ionic and transversal signals. The values for various experiments were determined through visual inspection of each time trace and the correlation of both measurement channels (electronic and ionic), as summarized in Tab. 6.3.

6.3 Machine Learning Models

6.3.1 Clustering Methods

In order to detect clusters, i.e. possible correlations, in the features extracted from all data, two-dimensional feature scatter plots are generated. The number of clusters k for all data sets and each combination of feature pairs were found on the basis of two internal clustering validation measures: the Silhouette SH [261] and the Calinski-Harabasz (CH) scores [262]. These are known statistical tools and quite distinct. The former is not very efficient in the case of very similar or sub-clusters [263], for which the latter provides more accurate results. Based on the clustering scores, the open-source k-means algorithm from the scikit-learn library [264] is applied. Given a specific number of clusters k, the algorithm is randomly initialized and iteratively determines the centroid positions based on the data position in the feature space [265]. In the field of nanopore sequencing, the application of the k-means algorithm can

enhance read accuracy by mitigating both systematic and random noise present in experimental datasets [266]. In order to have better control over the data clustering process, the standard k-means algorithm is opted for instead of utilizing an enhanced version that automatically determines the optimal number of clusters [265]. This decision is supported by the analysis employing two specific statistical scores and enables fine-tuning of the unsupervised learning scheme.

6.3.2 Classification Methods

In order to employ the learning process, especially for the supervised classification task, ML algorithms mostly from DL were used. The ML models applied here are widely used modern architectures fitting the purpose of the current work. Accordingly, XGB (referred to as XGBoost in the next), DNN, CNN, and LSTM RNN (referred to as LSTM in the next) were used as classifiers to compare their performance in predicting the molecule identity, i.e., in reducing errors in read-out. The scalable ML system for the XGBoost method is available as an open-source package [267]. For the different DL architectures used here, the implementation in the open-source DL library Keras [268] was chosen. This is a high-level API for building and training deep learning models that works with TensorFlow [269] running in the back-end. In the pre-processing pipeline, feature extraction is implemented using the scikit-learn library [264]. The features used to train the ML algorithms are extracted after the CUSUM algorithm is applied to the raw current traces, and the outliers have been removed. Whereas the learning data for the DNN and the CNN models are the feature-encoded current traces, for the baseline models using XGBoost and LSTM, the training data are the CUSUM level-fitted current traces. The data are normalized during the training process. In the case of LSTM, an additional module in the data pre-processing pipeline equalizes the length of the input data using masking and padding techniques. To optimize the topology of all these architectures and deal with overfitting, standard hyperparameter optimization practices are used. Further standard procedures, such as early stopping [192], are applied to reduce overfitting. Early stopping is applied in all learning models to stop training when the validation loss value diverges after a certain number of epochs that varies among the methods. Tuning the batch size can also control the fluctuations of the validation accuracy. To increase the speed of convergence and decrease overfitting, high values for the batch size are used, typically about half that of the training set. Overall, in the learning process, the training-validation-test dataset is split into the ratio 60 vs 20 vs 20% for nucleotide experiments and 80 vs 10 vs 10% for bio/molecules, due to the lack of data present in those experiments.

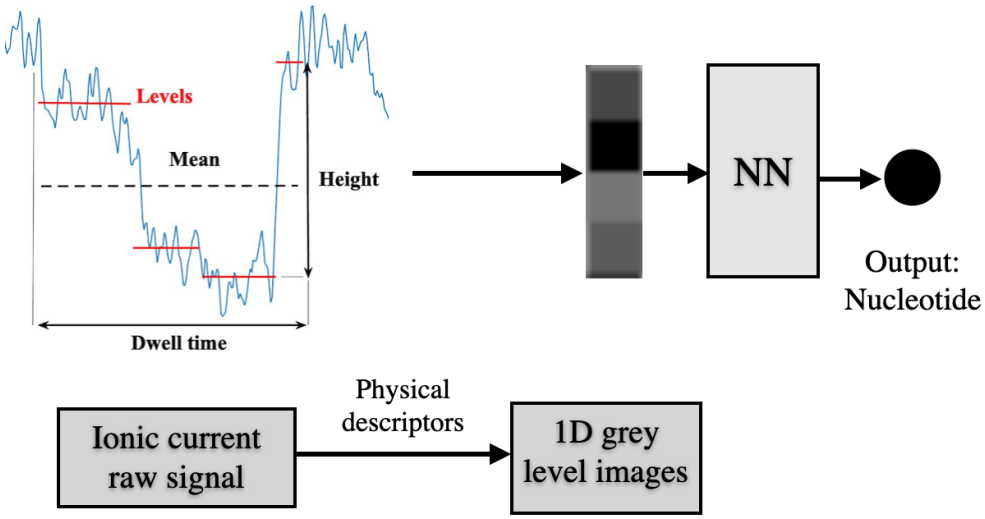


Fig. 6.3 The encoding process of mapping the ionic current raw signals, by means of the physical descriptors/features into grey-level scale images. The encoding is followed by the training procedure and the prediction of the nucleotide identity at the end of the pipeline. The images include 4 pixels corresponding to the four features for each single translocation experiment of a certain nucleotide [12].

Normalization: Gray-scale Images

It is well-established that CNN models excel in image processing [270]. To leverage this capability, both DNN and CNN models undergo training using four features mapped into 4×1 images. These images result from normalizing feature values to a grayscale range. This process, based on a pre-processing technique, generates images that assign distinct grayscale values to the four DNA nucleotides [271]. Specifically, four float values within the range [0-1] are assigned to different DNA nucleotides, transforming the DNA sequence into grayscale rows. Concatenating these rows produces 2D images, serving as input for the 2D CNN model. As the datasets correspond to translocation events of single nucleotides, each pixel doesn't represent a categorical value or nucleotide label but rather corresponds to numerical features. These features, obtained in the pre-processing of the two experiments, function as separate channels. The data are normalized in the range [0-255] in the grayscale. These 1D by 4 pixels images are used as input for NN architectures like fully connected DNN and CNN models. For the LSTM model and XGBoost, the raw signal of events from both experiments serves as input. The data transformation pipeline into grayscale images is illustrated in Fig. 6.3.

6.3.3 SHAP Analysis

In order to gain deeper insights into the clustering assessment, SHAP analysis [195] was applied to the features extracted from the nanopore data. This analysis enhances understanding of the decision-making process in supervised learning models. The resulting SHAP values offer valuable insights into how each feature contributes to distinguishing a specific observable or descriptor from the rest. This approach allows comprehension of the relative importance of features for each class independently. Such insights prove beneficial for feature selection, model interpretation, and the identification of specific characteristics that differentiate datasets or parts within datasets.

6.4 Results

In this results section, three key findings related to different dataset combinations are highlighted, as explained below:

- **Feature efficiency via clustering analysis:** In this subsection, various feature combinations for all datasets presented in Tab. 6.1 are analyzed, assessing their efficiency through clustering analysis. A total of thirteen datasets are included, encompassing d[X]MP nucleotides for Exp. A and Exp. B, ssDNA, dsDNA, and polyLys.
- **Nucleotide classification:** In this subsection, results for the classification of the two nucleotide analytes from Tab. 6.1 are presented across eight datasets, specifically d[X]MP nucleotides for Exp. A and Exp. B.
- **Feature importance via SHAP analysis:** In the third subsection, the feature assessment examines the impact on classification improvement by adding electronic features to ionic features using SHAP analysis. Here, the last three analytes in Tab. 6.1 (ssDNA, dsDNA, and polyLys) are utilized.

6.4.1 Feature Efficiency via Clustering Analysis

The primary phase of the analysis is initiated with an initial evaluation of feature relevance based on clustering. To accomplish this, various pair combinations of features are considered for each experiment. Specifically, ionic features for nucleotide translocations, ionic and electronic features for each detection channel separately, and both detection

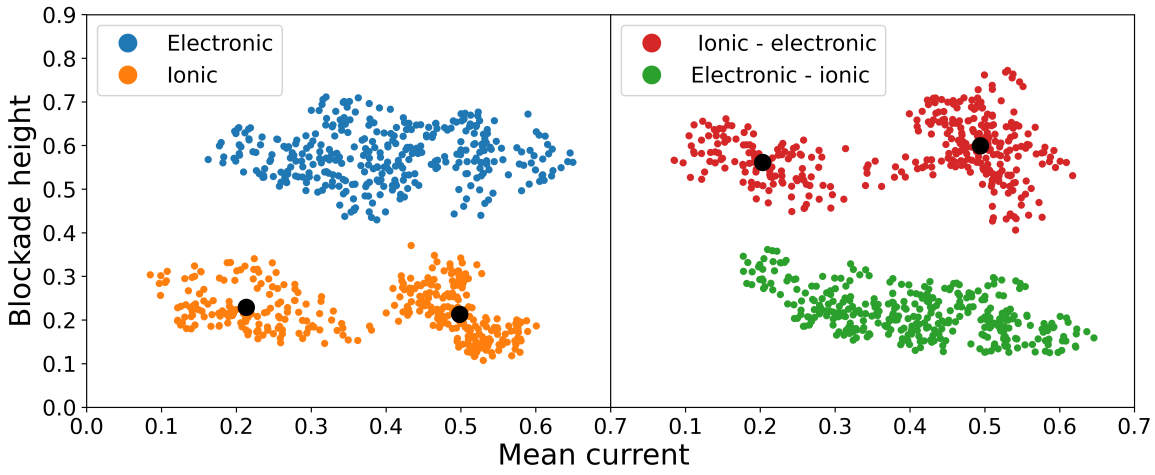


Fig. 6.4 Two dimensional graphs for the blockade mean and height features for the analyte 'ssDNA' from Tab. 6.1. The top panels represent the clusters for these features and the ionic (left) and electronic (right) measurement channels, as denoted by the legends. The lower panels evaluate both channels together: The black filled circles denote the center of each cluster.

channels concurrently for bio/molecule translocations. Pairwise clustering is performed for all features of the type ftr_i-ftr_j, where ftr denotes lvl, height, mean, dwell, drop, and i, j represent i and el (for ionic and electronic, respectively). Two-dimensional feature graphs are generated for all ftr_i-ftr_j pairs to identify clear clusters in the data. The similarity in patterns observed in previous clustering analyses in single nucleotide nanopore data [122] is not consistently reproduced here for all the data, primarily due to sparsity in some datasets. Nevertheless, even for sparser sets, distinct clusters seem to emerge for ftr_i-ftr_j combinations. To validate observations and determine the optimal number of clusters for all datasets, the S-score and CH-score are calculated, scanning across the number of clusters k found by k-means ranging from 2 to 10. Selecting dwell time as a feature yields an almost consistent S-score value as k increases for all experiments. On the other hand, the CH-score fails to identify the optimal number of clusters for a given data input, as its value diverges with k [272]. This discrepancy between the two scores underscores the limitation of using dwell time as a reliable feature for classification purposes. In contrast, when blockade height and mean current are chosen for the feature space, both scores exhibit a similar pattern of variation with respect to cluster size. The analysis indicates that using height together with the mean leads to well-defined and separated clusters in the feature space. Illustrative of this, in Fig. 6.4, the clustering of the feature space in relation to the current mean blockade height for the richest dataset ssDNA is depicted.

The results undergo normalization, achieved by scaling each feature between 0 and 1 using the min-max normalization function (MinMaxScaler) from scikit-learn [264]. This normalization aids in the convergence of k-means by accommodating the distinct range of ionic and electronic currents. For various channel combinations, whether ionic or electronic, two clearly distinguishable clusters are observed in the feature spaces of ionic-ionic and ionic-electronic, based on the highest score value calculated. This may imply the existence of two distinct (most probable) molecular configurations. Specifically, in the case of ssDNA, where the pore size (5.9 nm) exceeds the molecule's diameter (roughly 1 nm), the different clusters could be linked to single-file and folded translocation events. In the latter, the molecule translocates the pore in a folding conformation, leading to a more substantial blockade and deeper blockades. The importance of electronic current increases when examining clusters in the ionic-electronic feature space. Initially, the unsupervised scheme applied to ionic features reveals well-defined clusters. Further comparing the height for both channels (red points) shows an enhanced separation of clusters in the ionic-electronic case. These observations underscore the higher quality of cluster formation in the ionic-electronic feature space compared to the electronic one-channel space, with more distinct and well-separated clusters. This clarity and separation are further enhanced when utilizing height as a feature for both channels, particularly focusing on the height$_i$-height$_{el}$ feature space. This suggests a higher relevance of height over mean in defining clusters of distinct translocation events (folding, conformational changes, etc.). Consequently, both ionic and electronic current blockades exhibit signatures of such distinct events in their respective height values. While analyzing both channels separately can identify clusters, concurrent analysis of both channels using the height feature proves to be efficient for clear cluster distinction and information extraction.

6.4.2 Classification of Single Nucleotides

In the classification model propose here, the focus is on the efficiency of predicting the nucleotide identity (dAMP or dTMP or dCMP or dGMP). To achieve this, the classification performance of the different nucleotides is first discussed using the data from individual experiments (i.e. one pore diameter), and then the analysis proceeds by combining the data from different nanopore experiments (i.e. both pore diameters). In the end, comments are made on the connection of the classification schemes representative of unsupervised learning to the supervised models used in this work.

Method	Accuracy	F1-score
	Exp.A (3.3nm)	
CNN	0.88 ± 0.01	0.87 ± 0.01
DNN	0.86 ± 0.01	0.86 ± 0.01
LSTM	0.65 ± 0.22	0.64 ± 0.22
XGBoost	0.79 ± 0.02	0.78 ± 0.02
Most frequent	0.53	0.37
	Exp.B (2.8 nm)	
CNN	0.94 ± 0.01	0.93 ± 0.00
DNN	0.93 ± 0.00	0.92 ± 0.00
LSTM	0.81 ± 0.00	0.72 ± 0.00
XGBoost	0.95 ± 0.00	0.95 ± 0.00
Most frequent	0.81	0.72
	Exp.A+B	
CNN	0.92 ± 0.00	0.91 ± 0.00
DNN	0.88 ± 0.00	0.88 ± 0.01
LSTM	0.67 ± 0.04	0.60 ± 0.09
XGBoost	0.90 ± 0.01	0.90 ± 0.01
Most frequent	0.63	0.48

Table 6.4 Classification performance of the different ML algorithms based separately on the data from nucleotide experiments A (top results) and B (intermediate results), as well as from the combination of data from both Exp. A+B (bottom results). The pore diameters are also indicated. It should be noted that error values up to the second digit after the comma have been presented. For Exp. B, and thus also for some of the Exp. A+B results, the error is not 0, but in the order of 0.001. In order to keep it consistent with the other values in the table, this was rounded to zero [12].

Individual Experiments

The performance of the ML methods is first evaluated separately for each nanopore device, that is separate learning processes are carried out for the two experiments. The results on the accuracy obtained for each experiments are summarized in Tab. 6.4. All numbers are given with an accuracy up to the second decimal digit. As a first observation, the results from Exp. B are better, i.e. lead to a higher accuracy and learning scores. However, this could be expected, as part of the data (for dAMP and dCMP) in Exp. A are poorer than in Exp. B. Accordingly, the networks are trained with an overall more rich dataset n the case of Exp. B. In the case of the dGMP translocation data from Exp. A, the feature 'ionic blockade height' does not extract high correlations due to the longer dwell times of this dataset compared to the rest. As a result, due to its unclear cluster representation, the feature space does not allow

for the ML model to well fit the data [122]. This is, though, the only dataset where this situation was observed. Nevertheless, the CNN and DNN algorithms deal with low-correlated data more efficiently than the LSTM and XGBoost models.

For a further comparison of the learning methods, one should focus on the scores observed for a dummy classifier (see scikit-learn library) choosing always the most frequent class. It can be observed in the table that DNN and CNN models are the models that better adapt to the noise of the data and provide a better prediction. Nonetheless, the accuracy of every method decreases with an increase of the pore size. However, this could not strictly be related to the pore size, but partly also on the sparsity of the data for the larger pore, as mentioned above. For well-clustered datasets such as the ones in Exp. B, XGBoost and CNN are better alternatives than the ~~RNN~~ LSTM algorithm. However, CNN shows the highest prediction efficiency among all learning schemes even for datasets with very long dwell times such as dGMP from experiment A. In any case, for both experiments, the efficiency of the signal encoding through the data transformation proposed here is very promising. As a general remark, the performance of the DNN architecture for both nanopores is more efficient and accurate than LSTM, though it does not reach the robustness of our novel CNN method for the encoded time series.

Combination of Pore Sizes

Next, the data from both Exp. A and Exp. B are combined to train the ML models. The results are also summarized in Tab. 6.4. The most important observation is that the combination of all data leads overall to a very good prediction efficiency. This holds for all learning schemes apart from LSTM. In fact, the training and prediction efficiency are better than those from Exp. A alone and only slightly lower than those from Exp. B alone. This is a very interesting and promising result. The data used to train from both experiments are, of course, richer than those from the separate experiments. However, this alone does not justify the higher accuracy and the score. Similar types of data, i.e. the features from ionic current time series, but from different nanopores, have been combined. Overall, different pore sizes generate completely different configurations in the feature space for the same nucleotide type. Accordingly, the data combined from the two experiments do correspond to different values and clusters in the feature space. To this end, the fact that the accuracy, especially in the CNN training, is very close to that of Exp. B shows that combining distinct regions in the feature space, i.e. distinct values of the features, together with an optimum design/choice of the network can lead to very good predictions, i.e. enhance the read-out of nanopore data.

In order to further elucidate the differences among the results from all learning schemes, the confusion matrices for every scheme have been computed, as shown in Fig. 6.5. The figure reveals the influence, i.e. the quality of the specific datasets. Note that each dataset refers to the translocation of a certain nucleotide for both nanopores. That is, the data from both nanopores are considered for calculating the confusion coefficients for each nucleotide and each learning scheme. The figure shows the relation of the true identity (label) of a nucleotide compared to the predicted identity (label). A strong diagonal, that is high (blue) values on the diagonal of each panel, corresponds to a very good efficiency of the training model for all nucleotide datasets. This holds for DNN, CNN, and XGBoost. The LSTM model does not learn the sparser dGMP and dTMP datasets well. Accordingly, the other models (DNN, CNN, and XGBoost) can deal with data issues, such as dataset sparsity or large deviations in the values of the features. The latter could refer, for example, to long dwell times. The results in Tab. 6.4 and Fig. 6.5 underline a deep relation between clear cluster representations and the performance of the ML classifiers, that is unsupervised and supervised learning schemes.

The results from the combination of the two experiments reveal the ability of the CNN and DNN algorithms to adapt to the feature space even with low-correlated datasets, in the case of dGMP data from Exp. A combining experiments. Encoding techniques implicitly include pore size information through the physical features based on the well-defined and non-overlapping clusters observed previously in the feature space [122]. This provides flexibility in incorporating more pore sizes into the training, thereby up-scaling the model. Regarding the technical side of these results, dealing with outlier-prone datasets has been proven very demanding. Many concepts of data pre-processing and model refinement have to be applied to optimize the results. It could be shown that a physically motivated encoding of events can lead not only to good performance but also to a reduction in dimensionality. Regarding further optimization of the learning schemes alone, there is room for enhancing the efficiency of the XGBoost model. The LSTM model could be further refined using models capable of processing inputs with variable length and class weightings. The large number of parameters of the LSTM model makes it more dependent on a huge training dataset and more stiff, something that might hinder very good performance in the case of quite diverse data. On the other hand, simpler models like the XGBoost model may show better efficiency together with lower computational demand. Accordingly, future read-out algorithms for nanopore DNA sequencers could use encoding in traditional features together with the very promising 'ionic blockade height' to enhance their prediction possibilities. As

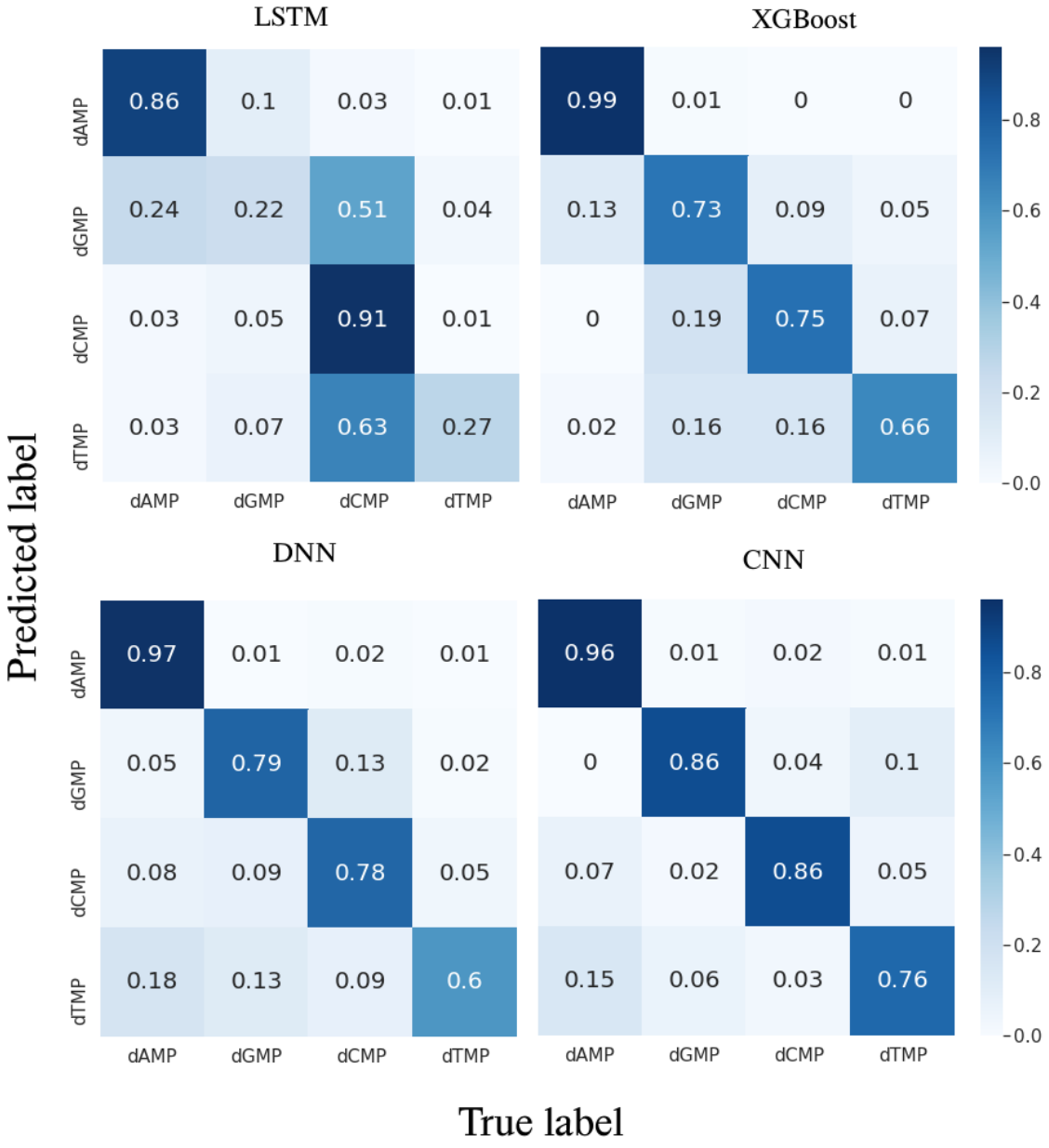


Fig. 6.5 Confusion matrices for the LSTM, XGBoost, DNN, and CNN models as denoted by the labels. All datasets from both experiments are represented. the 'True label' and 'Predicted label' refer to the true and predicted identity of the nucleotides [12].

a final and general remark, a strong improvement in the efficiency of supervised models encoding the data using unsupervised learning and data transformation techniques has been clearly demonstrated. Inspection of both the accuracy values in Tab. 6.4 and the confusion matrices in Fig. 6.5 can provide an indication of the minimum number of translocating DNA molecules needed for obtaining high base-calling accuracy from nanopores. The confusion matrices hint at higher accuracy with the XGBoost and the dAMP data ($\approx$ 4,000 events), while the next best results were obtained for dCMP ($\approx$ 1,100 events) and the CNN training. The accuracy for dAMP and XGBoost drops to 90% when all datasets are considered (see Tab. 6.4). This drop occurs due to the inclusion of the sparser datasets, especially that for dTMP ($\approx$ 370 events). Accordingly, translocation events in the range of 4,000 are expected to provide high base-calling accuracy, at least for single nucleotides. Note that reference is made here to the total dataset size and not the training set size.

6.4.3 Bimodal Feature Importance

In order to further assess the blockade features and the hierarchy of the information hidden in the bimodal data corresponding to ssDNA, dsDNA, and polyLys, the class balance related to the classification scheme XGBoost, applied on the data from all three experiments, is evaluated. According to the analyzed class balance, there is a large class/feature imbalance among the different datasets. The class balance corresponds to the amount of instances of each class, i.e. the features, present in the total dataset. In order to resolve this, both over- and under-sampling techniques have been applied, which remove data points from over-represented classes or create new artificial data points for under-represented classes in the training set. Specifically, a random over-sampler, which randomly duplicates data points from the existing ones for under-represented classes, and a random under-sampler, which randomly deletes data points from the over-represented classes, have been used [273, 274]. Note that, although for under-sampling the existing data points were duplicated and no new data points were generated, the training/testing and validation sets had to be split in advance to separate these completely. This procedure led to equally distributed classes with 300 data points in each experiment.

After addressing the data imbalance, the focus shifts to discerning a hierarchy within the extracted features. To serve as an indicator, the feature importance for both single and double channels is presented. Mean SHAP values for both ionic and electronic features extracted from 80nt ssDNA bimodal MoS2 translocations at 1M KCl and 1V are separately and collectively assessed. Accordingly, higher SHAP values

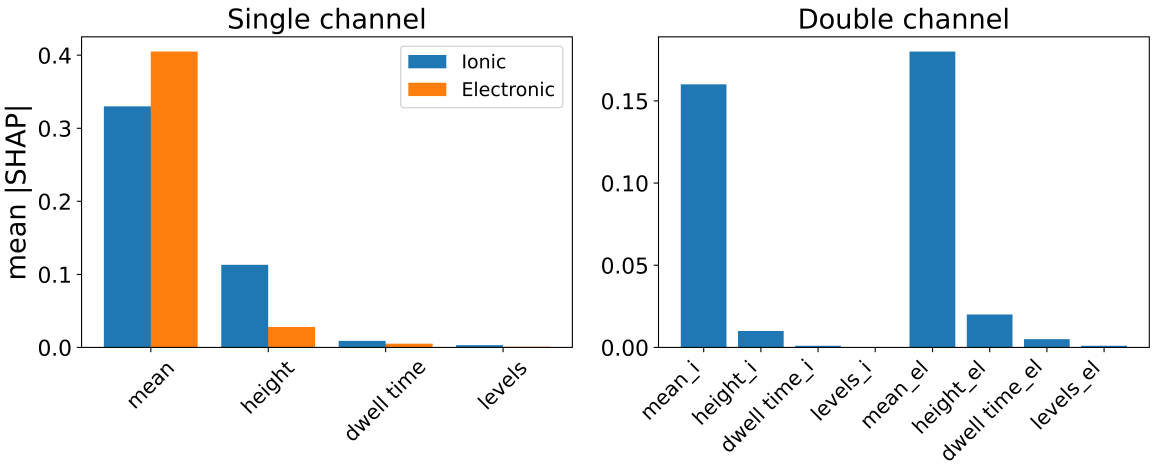


Fig. 6.6 The mean absolute SHAP values for the 80nt ssDNA dataset using the XGBoost classifier are depicted. On the left, a comparison of single-channel performance is shown, while on the right, the combination of both channels for molecule classification is presented.

indicate greater feature impact on the learning scheme, underlining that the respective feature/s are the most descriptive of the data and should be used in a training set for developing an accurate read-out protocol. In the figure, the SHAP value (or feature importance) is ranked from highest to lowest, in order to clearly visualize the feature hierarchy. In Fig. 6.6, the mean SHAP values for ssDNA using the XGBoost classifier are depicted. In the initial observation, the hierarchy obtained appears, on average, to be similar between the single and double channel cases. Significantly, the feature with the greater SHAP contribution is the mean, a trend observed consistently in both the ionic and electronic channels. In both channels, importance is attributed more to the mean and height, overshadowing the dwell time and level information in the current blockades. The latter, in certain cases, exhibit comparatively low or negligible importance, suggesting they lack essential information regarding the experimental details and molecular identity/sequence. The results clearly reveal that using both channels favors the bio/molecule detection, with the electronic mean feature turns out the one with the highest impact in the prediction.

6.5 Summary

In this chapter, ML has been applied to nanopore experimental data, specifically using current traces from nucleotide translocations through 2D MoS_2 nanopores. Unsupervised ML identified 'ionic blockade height' as a superior feature for nanopore

read-out, and its efficacy in predicting DNA nucleotides in a four-dimensional input space was demonstrated. CNN and XGBoost models outperformed LSTM, especially when transforming time series into 1D images. A pathway has been proposed that combines the strengths of ML with the scalability of DL models for diverse data types. Additionally, concurrent ionic and electronic current measurements from 2D MoS_2 nanopores threading biomolecules were analyzed. Clustering analysis emphasized the importance of the ionic-electronic feature space and the height feature in defining translocation events, with a strong correlation between clustering score and ML accuracy. The feature analysis and SHAP analysis revealed the significance of the mean current in defining blockades and translocating events. This information suggests a read-out protocol that focuses on learning from the mean and height features, enhancing sequencing accuracy for nanopores. Concurrently analysing both channels increases read-out efficiency and sequencing accuracy.

A step toward a general model for nanopore devices with different pore sizes is presented in this work, offering a faster alternative to LSTM models for next-generation nanopore DNA sequencers. By focusing on physically intuitive features, the aim is to explore the understanding of translocation events and signal compression. The applicability of the deep learning scheme extends beyond DNA nucleotides to other biomolecules and time series types, providing a valuable methodological alternative for error-free read-out of ultra-fast nanopore sequencers.

6.6 Acknowledgement

A substantial part of this chapter has been reproduced with permission from Refs. [12, 275], with the authorization of AIP Publishing and Nano Express. Gratitude is extended to the Radenovic group at EPFL, Switzerland, for their collaborative efforts and support.

Chapter 7

Learning the Critical Temperature of Superconductors through PDML

In this chapter, an innovative approach is introduced, combining ML techniques with *ab initio* descriptors to predict critical temperatures in novel superconductors. Unlike prior methods, the approach relies solely on descriptors derived from the electronic and atomic structure, enhancing interpretability and establishing a direct link between electronic orbitals and prediction accuracy. By selecting and extracting features related to electron concentration and electronegativity, state-of-the-art prediction accuracy is achieved using different learning models. Remarkably, this accuracy rivals more complex models with higher-dimensional feature spaces. Through careful dimensionality reduction, the pivotal role of the electron concentration for the different orbitals as a relevant descriptor is emphasized. These findings highlight the importance of these descriptors in predicting critical temperatures with a concise and physically meaningful feature space.

7.1 Datasets on Superconductors

In this work, a SuperCon-derived database [144, 276] was acquired by filtering out superconductors with incomplete compositions and repetitive formulas from the SuperCon database. Through this meticulous process, it was ensured that only superconductors with comprehensive compositions and unique formulas were included in the final dataset. By eliminating any redundancies or inconsistencies, a reliable and robust collection of superconducting data for further analysis and investigation was obtained, containing 12,340 superconductors.

7.2 Feature Selection

M. Pop et al. [277] demonstrated a correlation between the valence electron concentration and the critical temperature T_c of various inter-metallic superconductors. Building upon this concept, the electron concentration of different orbitals (s, p, d, and f) was initially utilized as a feature. To enhance this descriptor, it was further refined by calculating the electron concentrations (ECs) across distinct energy levels associated with each orbital type. In the end, 20 electron concentrations were collected as features. Inspired by the definition of VEC from Pop et al., the EC of a specific orbital i in a system with multiple species is given by:

$$EC_i = \frac{\sum_j m_j \cdot e_{ij}^S}{\sum_j m_j} \tag{7.1}$$

where:

- EC_i: Electron concentration in the i-th orbital.
- m_j: The count of atoms for the j-th species in the system.
- e_{ij}^S: The number of electrons in the i-th orbital for the j-th species.

Extra steps were taken to enhance the efficiency of the model. Outlier detection and feature removal were applied to streamline the inputs, focusing on the most relevant descriptors. Specifically, samples with zeros were excluded, and orbital features with both high correlation and low variance were pruned to refine the feature set.

7.3 PDML Model

The model proposed and applied for predicting the critical temperature of superconductors is based on a physics-driven selection and extraction of EC descriptors, which uniquely define the chemistry of the superconducting materials. These features are utilized to train an ML algorithm. For the latter, three different regressors were tested: (i) RF [278], (ii) XGB [267], and (iii) MLP [279]. Additionally, a linear regression model was employed for the learning, but it failed to accurately predict the critical temperature. In this work, the coefficient of determination R^2 was utilized as a metric for evaluating the predictive performance of the proposed models.

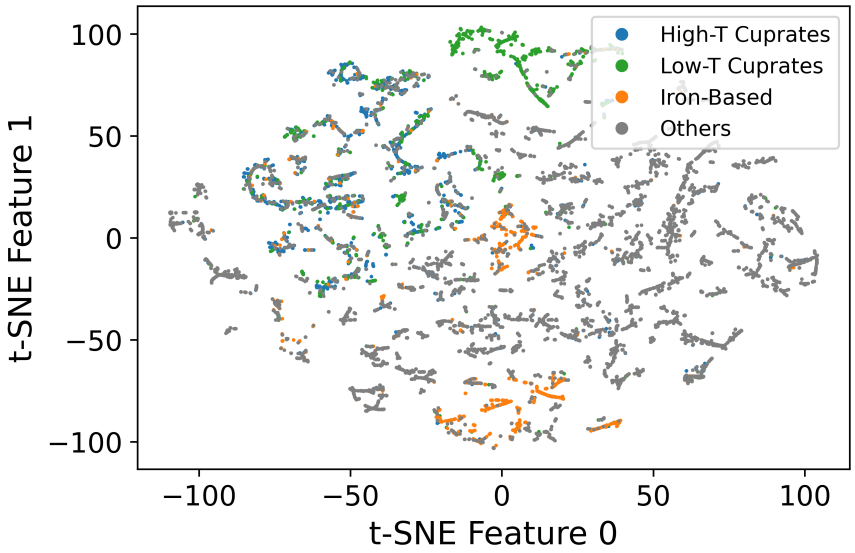


Fig. 7.1 t-SNE projection for the *EC* input space with 20 dimensions. The different groups iron-based, low-T_c and high-T_c cuprates, and 'others' are highlighted by the colors orange, green, blue and grey, respectively.

7.4 Results

7.4.1 Feature Efficiency via Dimensionality Reduction

The critical aspect of the learning process lies in the selection of features and the dimensionality of the feature space. To address this, an input space was crafted using only the electron concentration as input for the ML model. To gain insights into the 20-dimensional input space represented by electron concentrations (*EC*), the dimensionality reduction technique known as t-SNE (t-distributed Stochastic Neighbor Embedding) was employed. Developed by Laurens van der Maaten and Geoffrey Hinton in 2008 [280], t-SNE offers a powerful means of visualizing high-dimensional data. Fig. 7.1 illustrates the t-SNE projection, reducing the feature dimensionality from 20 to 2. The dataset was categorized into groups — iron-based, high-T_c cuprates, low-T_c cuprates, and others — to assess the separation in the projected space. The t-SNE algorithm adeptly distinguishes distinct groups and even subgroups within cuprate-based superconductors.

The subsequent step involved implementing and training the PDML model. In doing so, the *EC* feature space was integrated into three distinct ML regressors (i), (ii), and (iii), with the latter exhibiting the best performance overall. The train-test split is 85-15 %, randomly selected. Twenty repetitions of the learning model were

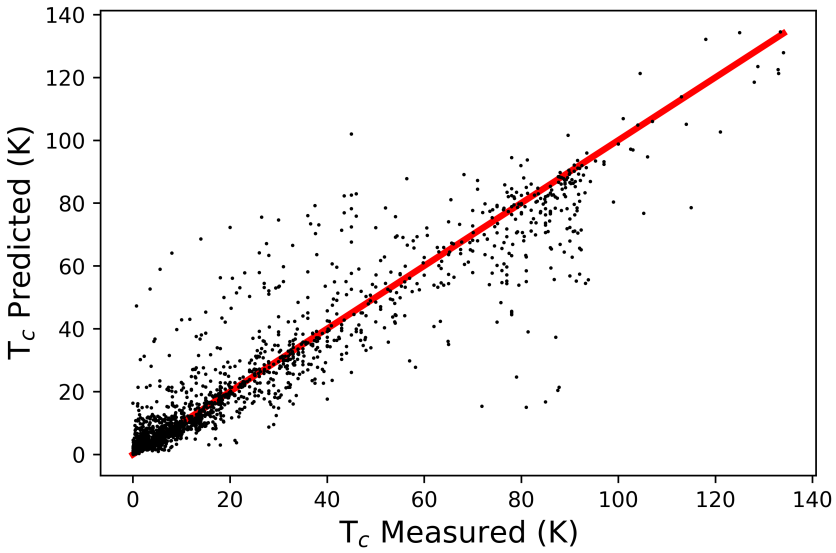


Fig. 7.2 Visual representation illustrating the comparison between measured and predicted T_c (K) values for the test set. The predictions were generated using MLP Regressor (iii).

performed to ensure statistical robustness. The results from these metrics reveal the deep connection between the level of electronic concentration and the T_c. This approach yielded an impressive average R^2 value of 0.878 ± 0.065 and a MAE (K) of 5.37 ± 0.143. The results derived from these metrics unveil a profound correlation between electronic concentration levels and T_c. This correlation is rooted in the simplicity of the definition and the reduction of feature space dimensionality, distinguishing it from other models presented in this paper and the existing literature. This accomplishment is particularly noteworthy considering the absence of property statistics and the exclusion of DL or clustering-derived features, which often sacrifice the physical interpretability of input variables. The significance of these results lies in the model's ability to provide accurate predictions while retaining the meaningfulness of input parameters, specifically the electron concentration by level and its connection to the critical temperature. These findings align with those presented in a referenced paper by [130], where the authors emphasize the limited correlation between thermodynamic features and the prediction of T_c across the spectrum of low and high T_c superconductors. The marginal improvement achieved by introducing derived features in other approaches from the literature [137, 128] does not significantly impact overall predictions or the categorization of superconductors. In Fig. 7.2 the test predictions of the PDML model are depicted. The model's performance is notably strong at both low and high temperatures. Never-

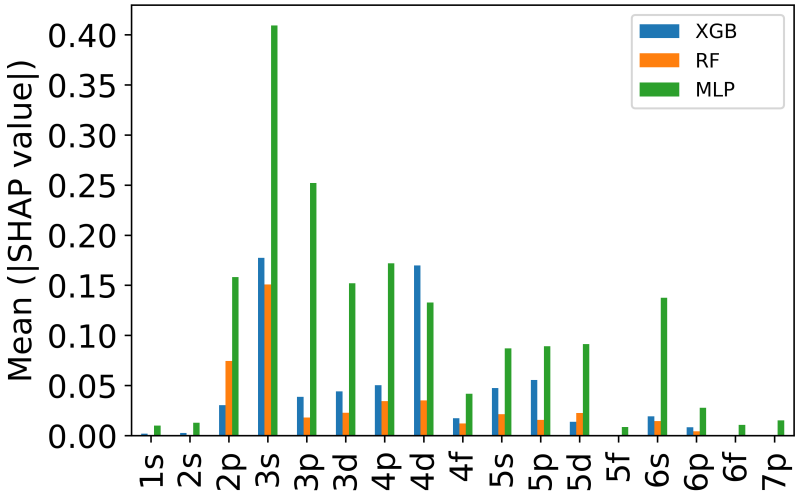


Fig. 7.3 Percentage of SHAP values contribution to the model for the different ECs.

theless, the sparse distribution in the mid-temperature region suggests a scarcity of superconductors at these specific T_c values.

7.4.2 Interpreting EC Features: SHAP Analysis

In order to assess which of the physical orbitals used are more relevant and strongly affect the prediction accuracy, the SHAP [195, 281, 282] analysis was employed to interpret the global model performance for the three different regressors (i), (ii), and (iii). By adopting this approach, valuable insights are gained into the significance of each feature in distinguishing a specific observable or descriptor from the others. This enables the comprehension of the relative importance of features for each class in an independent manner. In Fig 7.3, the feature importance for (i), (ii), and (iii) regressors applied to the PDML model is depicted. In view of the results, it can be observed that '3s' is the most influential descriptor, particularly in its vicinity. The feature removal eliminated the orbitals 4s, 6d, and 7s due to their low variance or high correlation when applied. Notably, MLP places more emphasis on the electronic concentration features compared to the tree-based regressor algorithms (XGB and RF). Interestingly, the feature importance values remain consistent across all the applied methods, enhancing the robustness of the hypothesis regarding the strong correlation between n_e and the T_c. This consistency supports the notion that n_e plays a pivotal role in influencing T_c predictions, regardless of the specific regression algorithm employed. This analysis underscores the novelty and significance of the model in terms of feature relevance, showcasing its potential and simplicity as a comprehensive surrogate for predicting

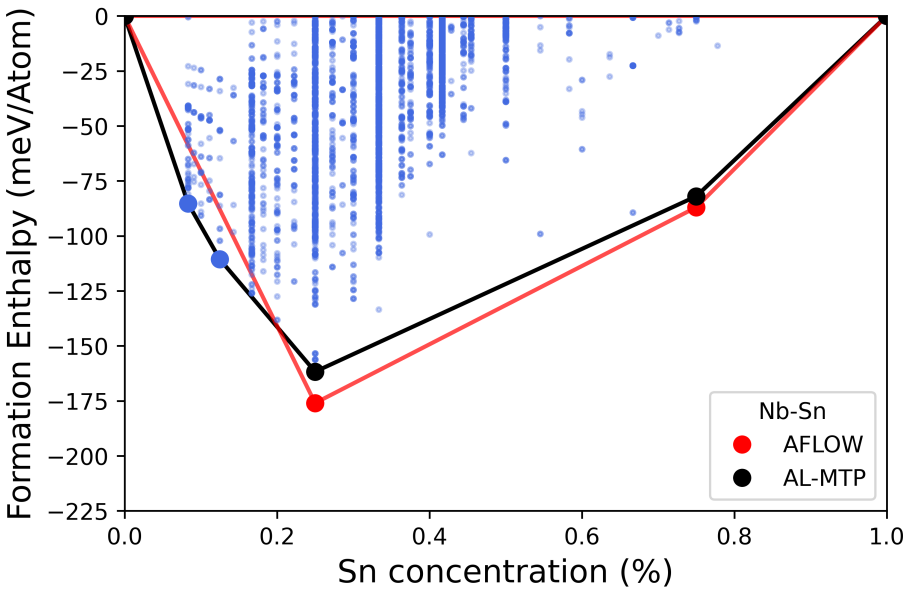


Fig. 7.4 Comparison between the convex hull obtained from AFLOW (represented by red dots and lines) and the one predicted by AL-MTP (depicted with black dots and lines). The algorithm identified two new stable structures (illustrated by blue dots).

T_c in superconductors. These findings strongly indicate that features derived from electronic properties are particularly prone to exhibiting correlations with the T_c of superconducting materials.

7.4.3 Prediction of Critical Temperature in HEA

As a test case, the novel PDML model with (iii) was employed to predict the T_c for a list of potentially new superconductors. To assess the model's performance, the convex hull of two binary systems containing A15 phases, including Nb-Sn and Nb-Al, was computed. The A15 phases form a series of intermetallic compounds defined by the chemical formula A_3B, wherein A signifies a transition metal, and B can be any element. Numerous compounds within this series exhibit Type-II superconductivity, making it a pivotal area of investigation given its myriad practical applications. In this instance, the pipeline outlined in Chapter 4 for AL-MTP was utilized to discover new stable phases and later predict their T_c. Fig. 7.4 illustrates the formation enthalpies of one of the A15 systems, Nb-Sn, including Nb_3Sn with the lowest formation enthalpy, and its convex hull calculated with AL-MTP (black line), compared to AFLOW (red line). For the Nb-Sn system, the algorithm identified two potentially stable phases: $Nb_{11}Sn$ and Nb_7Sn. In the Nb-Al system, a novel structure, $Nb_{11}Al$, was discovered by

Material	AFLOW	MLP
$Nb_{11}Sn$	7.46	9.47
Nb_7Sn	8.93	11.87
$Nb_{11}Al$	8.85	7.77

Table 7.1 Comparison of methods for T_c (K) prediction of novel binary phases with potential superconductivity.

AL-MTP. Subsequently, the T_c of the convex hull vertices were predicted in both cases. Following the early work of Matthias et al. on Nb_3Ge, which stated: "It is always the stoichiometric A15 compound which has the maximum transition temperature" [283], the expectation was that the new binary structures would exhibit lower T_c compared to the A15 Nb_3Sn phase. Tab. 7.1 illustrates the comparison between the method used in this work and AFLOW [79] for the known stable structures within the Nb-Sn system, as well as the novel $Nb_{11}Sn$ found by AL-MTP. It is evident that the T_c of the novel phases, characterized by higher formation enthalpies, is also lower than those of the A15 phases within their respective systems, namely Nb_3Sn (16.75 K) and Nb_3Al (18.8 K). In conclusion, the results suggest a linear correlation between formation enthalpy and T_c within the A-rich region of a system A-B featuring an A15 A_3B phase. Furthermore, the dimensionality reduction introduced by the novel approach enhances its uniqueness and potential for advancing research in the prediction of superconductivity.

7.5 Summary

In this work, a highly efficient PDML model for predicting the T_c of superconductors has been proposed. This model relies on the selective extraction of features based on the electronic and atomic properties of the elements forming the superconductors. The most important aspect of this model is the reduced dimensionality of the feature space, retaining the same accuracy as more involved schemes. Within this framework, the feature importance in the learning process was assessed, strongly emphasizing the high relevance of the electron concentration. In this regard, the proposed learning model is not a black box. Instead, it is rooted in a physics-driven selection of a small number of features that are strongly correlated with the electronic information inherent in the materials. Importantly, the *ab initio* definition of features stands out distinctly from the mean-approximation features used in similar learning schemes in the field. Therefore, the model presented here is specifically tailored for predicting a particular

signature or property of superconductors. The strength of this model lies in achieving a balance between accuracy and feature dimensionality, opening up possibilities for extension and modification to predict novel superconductor structures.

7.6 Acknowledgement

A substantial part of this chapter was reproduced with permission from Ref. [284], with the authorization of APL Materials.

Chapter 8

Conclusions

This thesis provides an overview of applications of physics-driven machine learning (PDML), where the inherent symmetries of physical systems were harnessed to enhance feature extraction and data representation techniques. The focus was on several specific domains, including copper-based alloys, nanopore detectors, and high-T_c superconductivity. Through this exploration, critical questions about the limitations of physics-inspired descriptors in ML, dimensionality reduction while maintaining prediction accuracy, the comparative performance of conventional ML versus physics-driven approaches, and the scalability of PDML in atomistic systems were addressed.

The investigation of copper-based alloys involved a combination of computer simulations and ML to uncover stable precipitate phases and study their mechanical properties. Quantum mechanical simulations, moment tensors and active learning (AL) were utilized to construct ML interatomic potentials, with a focus on moment tensor potentials (MTP) as utilized in this thesis. The objective was to identify potentially stable phases.

In the domain of nanopore detectors, ML models trained on experimental ionic blockade data were employed. The objective was to improve the efficiency of nucleotide identity read-out for error-free sequencing. Notably, the approach successfully reduced data dimensionality while preserving predictive accuracy, achieved by incorporating physical descriptors.

In the pursuit of predicting the superconducting critical temperature, the strategy emphasizes a physics-driven selection of a targeted set of features closely tied to material-specific electronic information. This approach ensures a balanced interplay between performance and feature dimensionality, facilitating the extension and customization of predictions for emerging superconductor structures. The methodology integrates insights from chapter 5 on copper-based alloys, amalgamating them with the AL-MTP

algorithm introduced in Section 5.1.2. Through the novel PDML regressor, the aim is to uncover potential new superconductors, with a particular focus on exploring both low- and high-T_c cuprates. This exploration aligns seamlessly with the primary research focus of the thesis, which centers around high-entropy copper-based alloys.

In the next phase of PDML development, physical models will be refined through the exploration of feature importance and the evaluation of ML model performance. The theoretical significance of physics-inspired descriptors will be analyzed, deepening the understanding and enhancing predictive capabilities of physical theories. Simultaneously, the optimization of PDML model performance will ensure reliable predictions, reinforcing its applicability across diverse scientific domains. This strategic approach will emphasize the synergy between physics and ML, offering an effective and computationally manageable pathway. Prioritizing these avenues will allow for the advancement of PDML without resorting to the complexities associated with techniques like equivariant models or physics constraints included in the loss functions. This marks a crucial step in the field's evolution toward breakthroughs and transformative applications in complex physical systems.

Through all these efforts, the gap between physics and machine learning was bridged, contributing to the emerging field of physics-driven machine learning. The work has the potential to revolutionize the understanding and modelling capabilities for complex physical systems, paving the way for transformative applications and advancements across various scientific disciplines.

Appendix A

Simple Classification of RNA Sequences of Respiratory-Related Coronaviruses

A simple, rapid, and efficient machine learning-based approach is proposed for analyzing and identifying respiratory-related virus sequences. The method, crucial for pandemic response, relies on genetic code rules and open reading frames. Using data from respiratory-related coronaviruses, features are extracted based on recurring nucleobase 3-tuples in RNA. The methodology involves counting nucleobase triplets, normalizing the count to sequence length, and applying principal component analysis (PCA). This triplet counting serves for classification, with potential for extending to DNA sequences from the herpes virus family. The scheme offers a fast, widely accessible, and portable detection method. While providing a foundational approach, it can be optimized and combined with supervised techniques for more accurate virus detection and sequencing. The relevance in identifying differences among similar viruses and their impact on biochemical analysis is discussed.

A.1 Data Collection and Preprocessing

A large amount of data has been collected for the RNA sequences of the respiratory-related coronaviruses family. The data refer to various viruses and were obtained from the NCBI database [285] and the Covid Predictor Project (CPP) [286], as summarized in Tab. A.1. In order to ensure that the sequence data did not only contain a protein sequence but the whole genome, the flag "complete sequences" has been used in the

GUI API of the databases. The virus data were stored in the FASTA format, which allows storing a large amount of genomic data in one file separated by a header line. To classify the data, we refer to the three viruses SARS-CoV, SARS-CoV-2, and MERS-CoV as the 'SARS/MERS viruses' and for the complete list of the respiratory-related coronaviruses in the table as the 'coronaviruses family.' For simplicity in the following, the notation 'SARS' is used for SARS-CoV and 'MERS' for MERS-CoV. Representative data from the herpes viruses family are also listed and will be used in the end for additional validation.

virus type	approx seq length	date accessed	# seqs	database
SARS-CoV	29751	16 Jun 2020	340	CPP [287]
SARS-CoV-2	29903	29 Sep 2020	22654	NCBI [285]
SARS-CoV-2 (PCA Set)	29903	21 Aug 2020	11118	NCBI
MERS-CoV	30111	07 Jul 2020	530	NCBI
Bovine Corona virus	31028	23 Sep 2020	309	NCBI
Camel Alphacoronavirus	27395	23 Sep 2020	70	NCBI
Duck coronavirus	27754	23 Sep 2020	425	NCBI
Alpha Herpesvirus	178101	05 Oct 2020	195	NCBI
Beta Herpesvirus	236100	05 Oct 2020	325	NCBI
Gamma Herpesvirus	172669	05 Oct 2020	657	NCBI

Table A.1 Types of viruses, approximate length of a virus genome sequence, date the data were accessed, number of complete virus genome sequences, and database for all RNA and DNA data used in the analysis.

A large number of virus sequences were processed for clustering and identification using the BIO python library [288]. The FASTA format was loaded using SeqIO to handle files containing multiple sequences. A bulk reading function was used to import the sequences into a list of dictionaries, capturing the name and sequence of each virus. The virus DNA list was then stored in a sequence object for further analysis. To extract information, the sequence was scanned for ORFs [289], which represent the regions containing protein information. A sliding window with a stride of three (for triplets) was employed to identify start and stop codons, as shown in Fig. A.1. The figure illustrates the ORF identification process within a nucleobase sequence using three different frames of reference (green, red, and blue). The green frame represents a complete ORF with a start codon 'ATG' and a stop codon 'TAA,' whereas the blue and red frames lack these codons. The labels in each color correspond to the amino acids encoded by the respective codons. Regarding the variability in triplet sequences, different frames of reference exist due to shifts of one or two nucleobases in the sequence, leading to distinct ORF starting points [290]. Additionally, the pair sequence, which

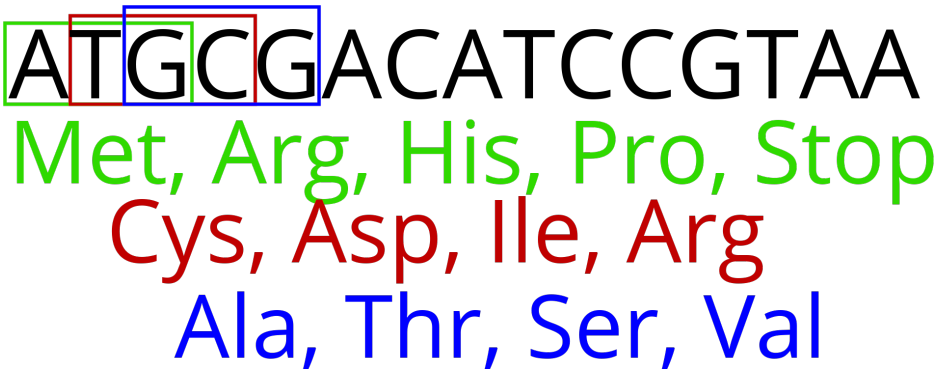


Fig. A.1 A sketch depicting the ORF identification process within a sequence of nucleobases (see text for more explanation). The labels in green, red, blue denote the amino acids ('Met' is methionine, 'Cys' is cysteine, etc.) that are made up from the respective codons.

contains the negative image of the original sequence, undergoes protein synthesis and expands the number of potential frames to six. In the analysis of positive-strand RNA viruses, which can be directly translated via protein synthesis, the focus was on analyzing three frames [290].

A.2 Feature Extraction

In order to extract feature vectors for the respiratory-related coronaviruses, the ORF1ab scheme was employed for longer sequences. Other virus types and families often have shorter ORFs, making length a distinguishing factor for the SARS virus family. Therefore, in feature extraction, the focus was on ORFs with a minimum length of 11,000 nucleobases. By sliding through the sequences with a stride of three, nucleobase triplets were isolated, as illustrated in Fig. A.2. The start codon ATG and stop codons TAG, TGA, and TAA were used for this analysis. These triplet counts were then utilized as features for subsequent clustering purposes. The nucleobase triplets in each ORF sequence are counted and normalized by the sequence length to create a feature vector. For data clustering, analysis, and information extraction, the scikit-learn library [264] was utilized. PCA was applied to reduce the dimensionality of the feature vector. This resulted in new features that are linear combinations of the triplet counts. The PCA transformed the feature vector into a two-dimensional feature space, preserving the variance of the original features. The triplet counts were taken into account, considering the codon degeneracy of the genetic code. In the analysis, the 'PCA Set' from Tab. A.1 was used for SARS-CoV-2. The resulting PCA matrix was

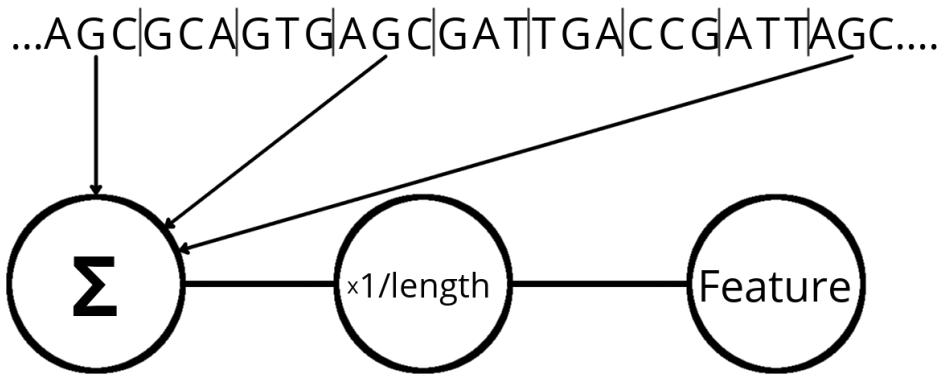


Fig. A.2 A sketch on the feature extraction scheme. Nucleobase triplets (the codons), shown on the top, are counted through the counter 'Σ' and normalized over the total length of the sequence to lead to each feature.

used to transform feature vectors from different virus types into the PCA feature space, leveraging data similarity. Consequently, the feature space represents a two-dimensional space formed by two feature vectors derived from PCA.

A.3 Implementation and Optimization

To ensure the accessibility and portability of the analysis tools across different computer architectures, Python was chosen as the programming language. Although Python may be slower compared to C, the cython library [291] was utilized to compile and accelerate the operations. Notably, the main sequence manipulation was achieved using a concise sliding window method, demonstrating the simplicity of the implementation. The feature vectors enable clustering analysis to identify and quantify separation among virus clusters in the feature space. For this purpose, the DBSCAN [292] and k-means [293] clustering schemes were utilized. To determine the optimal number of clusters and assess their accuracy, various clustering scores were employed, including SH [261], CH [262], Davies-Bouldin index (DB) [294], S_Dbw score [295], and SD_score [296]. While S_Dbw is often preferred for cluster identification [297], the analysis did not confirm this trend. Therefore, in cases where S_Dbw scores exhibited small differences or significant deviations among viruses, alternative clustering scores were employed. Notably, for the SARS-CoV-2 case, both relevant sets from Tab. A.1 were used for clustering, with only the older set used for building the PCA matrix. This enriched dataset incorporated updated releases, potentially including mutated sequences as discussed later. It is worth mentioning that all calculations were performed

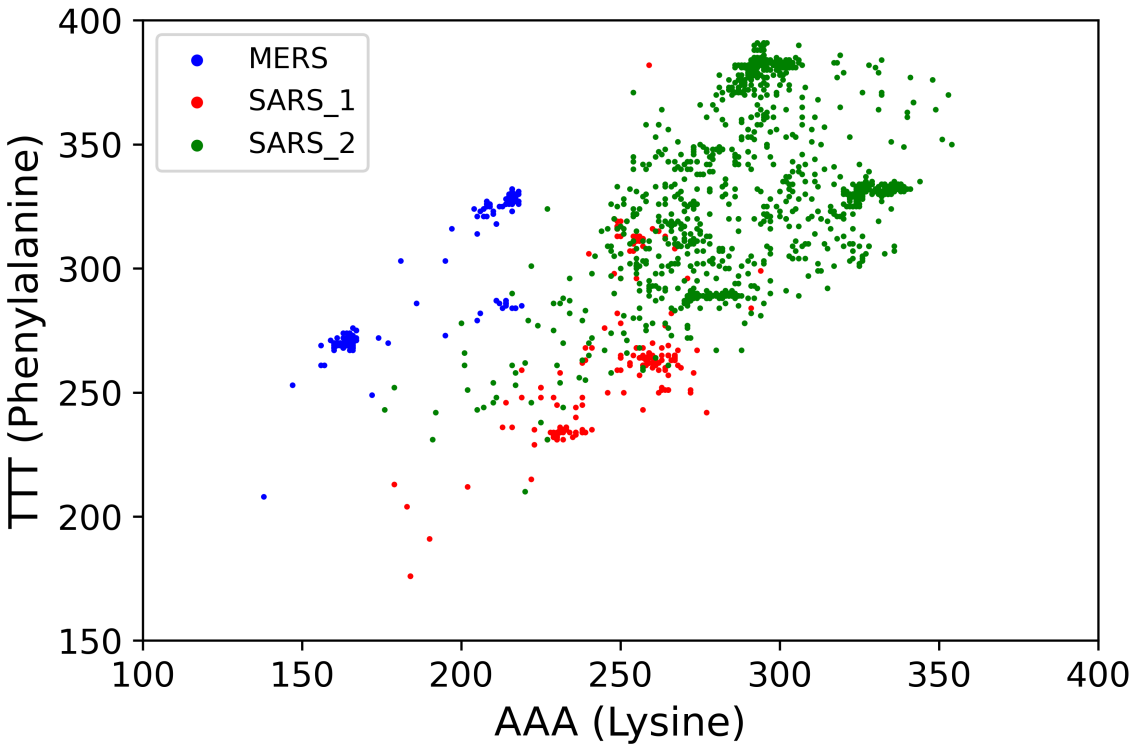


Fig. A.3 The feature space formed by two feature vectors ACC (Threonine) and GGC (Glycine) for the SARS/MERS virus family. The green, red, and blue symbols correspond to the SARS-CoV-2, SARS, and MERS viruses, respectively.

on a Raspberry Pi 4, taking approximately 30 minutes to process the entire set of coronaviruses, which is deemed acceptable given the dataset's size.

A.4 Clustering Analysis

The main findings are presented below, starting with the feature space analysis of the investigated coronaviruses. Following the methodology described earlier, features were extracted from ORFs with more than 11,000 nucleobases. The resulting feature spaces demonstrate a clear separation between different virus families. This is particularly evident in Fig. A.3, which depicts the feature space obtained from two randomly selected codons AAA and TTT, that corresponds to the amino acids lysine and phenylalanine, respectively. The promising results obtained from the clear and distinct clustering of SARS, SARS-CoV-2, and MERS viruses have led to the decision to expand the analysis to include all coronaviruses listed in Tab. A.1, applying PCA to the extracted PDML features. The clusters observed in the feature space for the

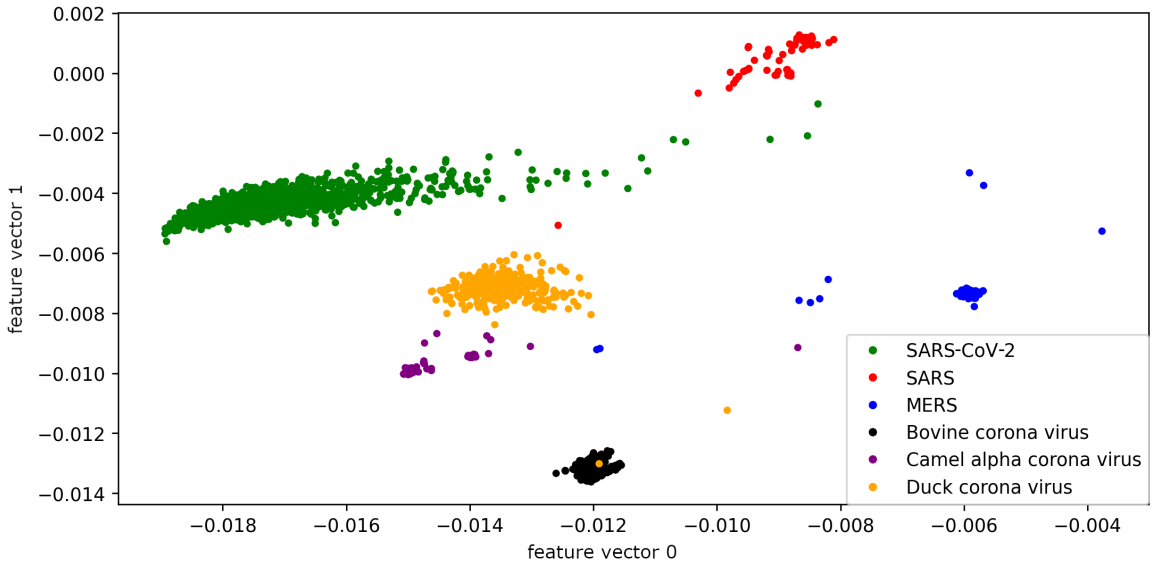


Fig. A.4 The feature space formed by two feature vectors ('0' and '1') from PCA for the corona virus family. The colors correspond to the different viruses as denoted by the legend.

coronaviruses are displayed in Fig. A.4. In most cases, well-defined clusters with dense centers and outliers further away are observed. However, a more dispersed distribution of features is exhibited by the Camel alpha coronavirus, which may be attributed to factors such as mutations over time or geographical variations. It is important to note that the outlier distribution for this virus is based on a smaller dataset compared to the larger datasets of SARS, MERS, and SARS-CoV-2. Additionally, the clusters for these viruses are relatively smaller due to their shorter sequences compared to SARS viruses. The clustering scores and data for these clusters are summarized in Tab. A.2.

To increase the diversity in the virus datasets, data from the herpes virus family have been included along with the previously used coronavirus family data. These additional data correspond to the last three entries in Tab. A.1 and serve as a validation of the approach in distinguishing viruses and their families. The clustering results in the feature space are depicted in Fig. A.5. It is evident that the clusters representing the herpes family are well separated from the others in the PCA feature space. However, it should be noted that only the gamma herpes virus is visible in this feature space, as it is the only one with an ORF exceeding 11,000 nucleobases in its positive sense frames. The results of the clustering analysis indicate that more data are necessary to achieve a clear separation among different coronaviruses in the PCA feature space. Nonetheless, distinct clusters can still be observed for most cases with the available data. In terms

cluster	SH	CH	DB	S_Dbw	SD	eps value
DBSCAN						
2	0.899	15961.958	**0.535**	0.750	**5.136**	0.1900
3	0.919	15077.836	1.353	0.559	7.591	0.1500
4	0.948	46796.690	0.811	0.407	13.382	0.0900
5	**0.954**	**162956.426**	0.776	0.401	14.003	0.0800
6	0.951	145818.938	0.781	0.374	14.351	0.0400
8	0.951	119730.571	0.731	**0.261**	33.030	0.0200
12	0.914	62971.981	1.087	0.289	105.356	0.0100
k-means						
2	0.927	50233.283	0.626	1.932	**4.121**	
3	0.940	62878.163	0.727	1.331	5.377	
4	0.954	109480.204	0.395	0.547	4.341	
5	**0.955**	231772.508	**0.205**	0.276	5.931	
6	0.926	386880.318	0.289	0.263	15.574	
7	0.904	453096.583	0.370	0.246	29.796	
8	0.904	543059.277	0.372	**0.221**	30.940	
9	0.893	**601798.301**	0.408	0.230	41.851	

Table A.2 Clustering scores obtained with DBSCAN (top) and k-means (bottom) for the set of the corona virus family feature vectors. The bold number in the first column ('clusters') indicates the expected number of resulting clusters. The bold numbers in the other columns emphasize the best scoring result. The eps value in the last column (top results) denotes the value at which the DBSCAN clustering was performed.

of the clustering scores, the S_Dbw score proved to be inefficient in determining the appropriate number of clusters for the virus families considered here. While other clustering scores exhibited variations around the expected number of clusters, the S_Dbw score consistently indicated a larger number of clusters. In this feature space distribution, DBSCAN outperformed k-means, as evidenced by the clustering scores that closely aligned with the expected number of clusters for DBSCAN.

Fig. A.5 visually demonstrates a clear separation among different virus types, with minimal mixing observed in the feature space. The spread of the SARS-CoV-2 virus from the dense cluster center to a more sparse distribution is potentially indicative of mutations in the ORF1ab since its initial discovery (refer to Tab. A.1 for data access dates). The MERS virus shows a broad distribution in the feature space, likely due to mutations and other variations. The bovine and duck coronaviruses exhibit excellent clustering, while the limited data availability for the camel alpha coronavirus prevents determining its cluster shape, though it is situated near the duck coronavirus cluster.

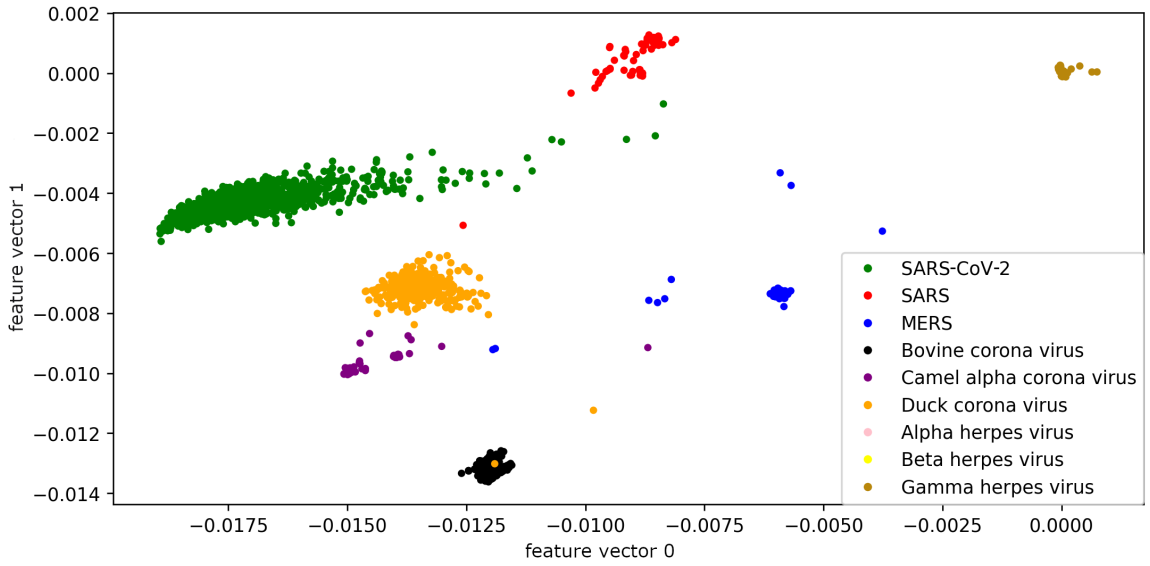


Fig. A.5 The feature space formed by two feature vectors ('0' and '1') from PCA for the corona and the herpes virus families. The colors correspond to the different viruses as denoted by the legend.

A.5 Conclusions

The findings suggest that there is no overlap in the feature space between the shorter herpes ORFs and the coronavirus family, offering potential for further research in developing a comprehensive virus classifier. However, to draw conclusive results, an extensive analysis of other virus families is required. This study primarily serves as a proof-of-concept for efficiently identifying virus clusters, focusing on the coronavirus family. This model highlights the crucial importance of tapping into the hidden biological information within ORFs for effective analysis of biological data. The proposed biology-driven analysis scheme demonstrates high efficiency in identifying and distinguishing viruses. Moreover, this approach is portable across different architectures, enabling fast and on-the-fly detection, thus making it widely accessible.

A.6 Acknowledgement

Most part of this chapter was reproduced from Ref [298] with the permission of ACS Omega.

References

[1] T. E. Smidt, "Euclidean symmetry and equivariance in machine learning," *Trends in Chemistry*, vol. 3, no. 2, pp. 82–85, 2021, special Issue: Machine Learning for Molecules and Materials. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589597420302641

[2] B. Qian, J. Su, Z. Wen, D. N. Jha, Y. Li, Y. Guan, D. Puthal, P. James, R. Yang, A. Y. Zomaya *et al.*, "Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey," 2020.

[3] "GitHub - NSHipster/DBSCAN: Density-based spatial clustering of applications with noise — github.com," https://github.com/NSHipster/DBSCAN, [Accessed 04-10-2023].

[4] M. Taboga, "Decision tree, lectures on machine learning," https://www.statlect.com/machine-learning/decision-tree, [Accessed 18-10-2023].

[5] C. C. Aggarwal, *Neural Networks and Deep Learning*. Cham: Springer, 2018.

[6] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, Jun. 2021. [Online]. Available: https://doi.org/10.1038/s42254-021-00314-5

[7] A. Bogatskiy *et al.*, "Symmetry Group Equivariant Architectures for Physics," in *Snowmass 2021*, 3 2022.

[8] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, "Performance and cost assessment of machine learning interatomic potentials," *J Phys Chem A*, vol. 124, no. 4, pp. 731–745, Jan. 2020.

[9] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," *Computational Materials Science*, vol. 156, pp. 148–156, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025618306372

[10] L. Lin, J. Lu, and L. Ying, "Numerical methods for Kohn–Sham density functional theory," *Acta Numerica*, vol. 28, p. 405–539, 2019.

[11] Á. Díaz Carral, X. Xu, S. Gravelle, A. YazdanYar, S. Schmauder, and M. Fyta, "Stability of binary precipitates in Cu-Ni-Si-Cr alloys investigated through active learning," *Materials Chemistry and Physics*, vol. 306, p. 128053, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0254058423007617

[12] Á. Díaz Carral, M. Ostertag, and M. Fyta, "Deep learning for nanopore ionic current blockades," *The Journal of Chemical Physics*, vol. 154, no. 4, p. 044111, 01 2021. [Online]. Available: https://doi.org/10.1063/5.0037938

[13] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021. [Online]. Available: https://doi.org/10.1007/s12525-021-00475-2

[14] J. Pateras, P. Rana, and P. Ghosh, "A taxonomic survey of physics-informed machine learning," *Applied Sciences*, vol. 13, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/12/6892

[15] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 614–633, 2023.

[16] D. Li, Y. Xu, M. Zhao, J. Zhu, and S. Zhang, "Knowledge-driven machine learning and applications in wireless communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 454–467, 2022.

[17] T. Dash, S. Chitlangia, A. Ahuja, and A. Srinivasan, "A review of some techniques for inclusion of domain-knowledge into deep neural networks," *Scientific Reports*, vol. 12, no. 1, p. 1040, Jan. 2022. [Online]. Available: https://doi.org/10.1038/s41598-021-04590-0

[18] B. G. Humm, P. Archer, H. Bense, C. Bernier, C. Goetz, T. Hoppe, F. Schumann, M. Siegel, R. Wenning, and A. Zender, "New directions for applied knowledge-based AI and machine learning," *Informatik Spektrum*, vol. 46, no. 2, pp. 65–78, Apr. 2023. [Online]. Available: https://doi.org/10.1007/s00287-022-01513-9

[19] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nature Communications*, vol. 13, no. 1, p. 2453, May 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-29939-5

[20] M. Dax, S. R. Green, J. Gair, M. Deistler, B. Schölkopf, and J. H. Macke, "Group equivariant neural posterior estimation," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=u6s8dSporO8

[21] P. L. Rossiter, *The Electrical Resistivity of Metals and Alloys*, ser. Cambridge Solid State Science Series. Cambridge University Press, 1987.

[22] C. Wang, H. Fu, L. Jiang, D. Xue, and J. Xie, "A property-oriented design strategy for high performance copper alloys via machine learning," *npj Computational Materials*, vol. 5, no. 1, p. 87, 2019. [Online]. Available: https://doi.org/10.1038/s41524-019-0227-7

[23] R. Pohja, H. Vestman, P. Jauhiainen, and H. Haenninen, "Narrow gap arc welding experiments of thick copper sections," Finland, Tech. Rep. POSIVA–03-09, 2003. [Online]. Available: http://inis.iaea.org/search/search.aspx?orig_q=RN:35052321

[24] X. Cui, Y. Wu, G. Zhang, Y. Liu, and X. Liu, "Study on the improvement of electrical conductivity and mechanical properties of low alloying electrical aluminum alloys," *Composites Part B: Engineering*, vol. 110, pp. 381–387, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359836816306679

[25] H. S. Abdo, A. H. Seikh, J. A. Mohammed, and M. S. Soliman, "Alloying elements effects on electrical conductivity and mechanical properties of newly fabricated Al based alloys produced by conventional casting process," *Materials*, vol. 14, no. 14, p. 3971, Jul 2021. [Online]. Available: http://dx.doi.org/10.3390/ma14143971

[26] F. C. Maier, S. Hocker, S. Schmauder, and M. Fyta, "Interplay of structural, electronic, and transport features in copper alloys," *Journal of Alloys and Compounds*, vol. 777, pp. 619–626, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838818340398

[27] H. Sehitoglu, T. Foglesong, and H. J. Maier, "Precipitate effects on the mechanical behavior of aluminum copper alloys: Part i. experiments," *Metallurgical and Materials Transactions A*, vol. 36, no. 13, pp. 749–761, 2005. [Online]. Available: https://doi.org/10.1007/s11661-005-1006-2

[28] J. Li, H. Chen, Q. Fang, C. Jiang, Y. Liu, and P. K. Liaw, "Unraveling the dislocation–precipitate interactions in high-entropy alloys," *International Journal of Plasticity*, vol. 133, p. 102819, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749641920302011

[29] S. Peng, Y. Wei, and H. Gao, "Nanoscale precipitates as sustainable dislocation sources for enhanced ductility and high strength," *Proceedings of the National Academy of Sciences*, vol. 117, no. 10, pp. 5204–5209, 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1914615117

[30] T. Fujii, H. Nakazawa, M. Kato, and U. Dahmen, "Crystallography and morphology of nanosized Cr particles in a Cu–0.2% Cr alloy," *Acta Materialia*, vol. 48, no. 5, pp. 1033–1045, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359645499004115

[31] Y. Jin, K. Adachi, T. Takeuchi, and H. G. Suzuki, "Ageing characteristics of Cu–Cr in-situ composite," *Journal of Materials Science*, vol. 33, pp. 1333–1341, 1998.

[32] S. A. Lockyer and F. W. Noble, "Precipitate structure in a Cu-Ni-Si alloy," *Journal of Materials Science*, vol. 29, no. 1, pp. 218–226, 1994. [Online]. Available: https://doi.org/10.1007/BF00356596

[33] Q. Lei, Z. Li, T. Xiao, Y. Pang, Z. Xiang, W. Qiu, and Z. Xiao, "A new ultrahigh strength Cu–Ni–Si alloy," *Intermetallics*, vol. 42, pp. 77–84, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966979513001593

[34] W. Wang, H. Kang, Z. Chen, Z. Chen, C. Zou, R. Li, G. Yin, and T. Wang, "Effects of Cr and Zr additions on microstructure and properties of Cu-Ni-Si alloys," *Materials Science and Engineering: A*, vol. 673, pp. 378–390, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921509316307754

[35] O. Samoilova and E. Trofimov, "Phase equilibria in the copper-rich corner of the Cu-Ni-Si-Cr system," *Materials Science Forum*, vol. 870, pp. 107–112, 09 2016.

[36] N. T. Kareva, I. L. Yakovleva, and O. V. Samoilova, "On the precipitation strengthening of $Cu_{2.6}Ni_{0.6}Si_{0.6}Cr$ bronzes," *Physics of Metals and Metallography*, vol. 118, no. 8, pp. 795–801, 2017. [Online]. Available: https://doi.org/10.1134/S0031918X17080075

[37] K. Sufryd, N. Ponweiser, P. Riani, K. W. Richter, and G. Cacciamani, "Experimental investigation of the Cu–Si phase diagram at x(Cu)>0.72," *Intermetallics*, vol. 19, no. 10, pp. 1479–1488, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966979511001634

[38] J. Cheng, B. Tang, F. Yu, and B. Shen, "Evaluation of nanoscaled precipitates in a Cu–Ni–Si–Cr alloy during aging," *Journal of Alloys and Compounds*, vol. 614, pp. 189–195, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838814014455

[39] R. R. Chromik, W. K. Neils, and E. J. Cotts, "Thermodynamic and kinetic study of solid state reactions in the Cu-Si system," *Journal of Applied Physics*, vol. 86, no. 8, pp. 4273–4281, 1999. [Online]. Available: https://doi.org/10.1063/1.371357

[40] E. Dodony, G. Z. Radnóczi, and I. Dódony, "Low temperature formation of copper rich silicides," *Intermetallics*, vol. 107, pp. 108–115, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966979518310987

[41] C. Y. Wen and F. Spaepen, "In situ electron microscopy of the phases of Cu_3Si," *Philosophical Magazine*, vol. 87, no. 35, pp. 5581–5599, 12 2007. [Online]. Available: https://doi.org/10.1080/14786430701675829

[42] B. Polat, O. Eryilmaz, O. Keleş, A. Erdemir, and K. Amine, "Compositionally graded SiCu thin film anode by magnetron sputtering for lithium ion battery," *Thin Solid Films*, vol. 596, pp. 190–197, 2015, the 42nd International Conference on Metallurgical Coatings and Thin Films. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040609015010378

[43] S. B. Lee, D.-K. Choi, F. Phillipp, K.-S. Jeon, and C. K. Kim, "In situ high-resolution transmission electron microscopy study of interfacial reactions of Cu thin films on amorphous silicon," *Applied Physics Letters*, vol. 88, no. 8, p. 083117, 2006. [Online]. Available: https://doi.org/10.1063/1.2179143

[44] T. Buonassisi, M. A. Marcus, A. A. Istratov, M. Heuer, T. F. Ciszek, B. Lai, Z. Cai, and E. R. Weber, "Analysis of copper-rich precipitates in silicon: Chemical state, gettering, and impact on multicrystalline silicon solar cell material," *Journal of Applied Physics*, vol. 97, no. 6, p. 063503, 2005. [Online]. Available: https://doi.org/10.1063/1.1827913

[45] P. Eckerlin and H. Kandler, "Structure data of elements and intermetallic phases · Cu-Se - Dy-Ge: Datasheet from landolt-börnstein - group iii condensed matter · volume 6: "structure data of elements and intermetallic phases" in springermaterials (https://doi.org/10.1007/10201454_41)," accessed 2023-01-17. [Online]. Available: https://materials.springer.com/lb/docs/sm_lbs_978-3-540-36859-5_41

[46] N. Mattern, R. Seyrich, L. Wilde, C. Baehtz, M. Knapp, and J. Acker, "Phase formation of rapidly quenched Cu-Si alloys," *Journal of Alloys and Compounds*, vol. 429, no. 1, pp. 211–215, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838806004609

[47] X. Yan and Y. Chang, "A thermodynamic analysis of the Cu-Si system," *Journal of Alloys and Compounds*, vol. 308, no. 1, pp. 221–229, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092583880000983X

[48] A. Dahal, J. Gunasekera, L. Harringer, D. K. Singh, and D. J. Singh, "Metallic nickel silicides: Experiments and theory for Ni-Si and first principles calculations for other phases," *Journal of Alloys and Compounds*, vol. 672, pp. 110–116, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838816304005

[49] P. Nash and A. Nash, "The Ni-Si (nickel-silicon) system," *Bulletin of Alloy Phase Diagrams*, vol. 8, no. 1, pp. 6–14, 1987. [Online]. Available: https://doi.org/10.1007/BF02868885

[50] P. Eckerlin and H. Kandler, "Structure data of elements and intermetallic phases · Ni-Se-V - Os-Pu: Datasheet from landolt-börnstein - group iii condensed matter · volume 6: "structure data of elements and intermetallic phases" in springermaterials (https://doi.org/10.1007/10201454_59)," accessed 2023-01-19. [Online]. Available: https://materials.springer.com/lb/docs/sm_lbs_978-3-540-36859-5_59

[51] C. Watanabe and R. Monzen, "Coarsening of δ-Ni_2Si precipitates in a Cu-Ni-Si alloy," *Journal of Materials Science*, vol. 46, no. 12, pp. 4327–4335, 2011. [Online]. Available: https://doi.org/10.1007/s10853-011-5261-x

[52] D. Connétable and O. Thomas, "First-principles study of nickel-silicides ordered phases," *Journal of Alloys and Compounds*, vol. 509, no. 6, pp. 2639–2644,

2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838810026472

[53] B. Ren, D.-H. Lu, R. Zhou, D.-P. Ji, M.-W. Hu, and J. Feng, "First principles study of stability, mechanical, and electronic properties of chromium silicides," *Chinese Physics B*, vol. 27, 2018.

[54] E. Mazzega, M. Michelini, and F. Nava, "Electrical properties of chromium silicide films: Cr_3Si and Cr_5Si_3," *Journal of Physics F: Metal Physics*, vol. 17, no. 5, p. 1135, May 1987. [Online]. Available: https://dx.doi.org/10.1088/0305-4608/17/5/013

[55] T. Dasgupta, J. Etourneau, B. Chevalier, S. F. Matar, and A. M. Umarji, "Structural, thermal, and electrical properties of $CrSi_2$," *Journal of Applied Physics*, vol. 103, no. 11, p. 113516, 2008. [Online]. Available: https://doi.org/10.1063/1.2917347

[56] L. F. Mattheiss, "Calculated structural properties of $CrSi_2$, $MoSi_2$, and WSi_2," *Phys. Rev. B*, vol. 45, pp. 3252–3259, Feb 1992. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.45.3252

[57] S. Cui and I.-H. Jung, "Thermodynamic assessments of the Cr-Si and Al-Cr-Si systems," *Journal of Alloys and Compounds*, vol. 708, pp. 887–902, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838817308113

[58] K. S. Chan, Y. D. Lee, and Y. M. Pan, "First-principles computations of mechanical properties of Ni_2Cr and Ni_2Mo," *Metallurgical and Materials Transactions A*, vol. 37, no. 3, pp. 523–537, 2006. [Online]. Available: https://doi.org/10.1007/s11661-006-0024-z

[59] D. Alontseva and A. Russakova, "The structure-phase compositions and mechanical properties of Ni-Cr-Al-based alloy after strong deformation and low-temperature aging," *Advanced Materials Research*, vol. 875-877, pp. 558–561, 02 2014.

[60] P. E. A. Turchi, L. Kaufman, and Z.-K. Liu, "Modeling of Ni-Cr-Mo based alloys: Part i—phase stability," *Calphad*, vol. 30, no. 1, pp. 70–87, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0364591605000970

[61] N. Aerne, D. J. Sprouster, and J. D. Tucker, "The formation and evolution of Ni_2Cr precipitates in Ni–Cr model alloys as a function of stoichiometry characterized by synchrotron x-ray diffraction," *Materials Science and Engineering: A*, vol. 856, p. 143930, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921509322013090

[62] J. Teeriniemi, P. Taskinen, and K. Laasonen, "First-principles investigation of the Cu–Ni, Cu–Pd, and Ni–Pd binary alloy systems," *Intermetallics*, vol. 57, pp. 41–50, 2015.

[63] T. Muroga, H. Watanabe, N. Yoshida, H. Kurishita, and M. Hamilton, "Microstructure and tensile properties of neutron irradiated Cu and Cu_5Ni containing isotopically controlled boron," *Journal of Nuclear Materials*, vol. 225, pp. 137–145, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022311595000275

[64] J. A. Mary, A. Manikandan, L. J. Kennedy, M. Bououdina, R. Sundaram, and J. J. Vijaya, "Structure and magnetic properties of Cu-Ni alloy nanoparticles prepared by rapid microwave combustion method," *Transactions of Nonferrous Metals Society of China*, vol. 24, no. 5, pp. 1467–1473, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1003632614632143

[65] A. Pasturel, V. Drchal, J. Kudrnovský, and P. Weinberger, "First-principles study of surface segregation in Cu-Ni alloys," *Phys. Rev. B*, vol. 48, pp. 2704–2710, Jul 1993. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.48.2704

[66] T. Li, C. He, W. Zhang, and M. Cheng, "Structural and melting properties of Cu-Ni clusters: A simulation study," *Journal of Alloys and Compounds*, vol. 752, pp. 76–84, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838818314464

[67] M. A. Turchanin, P. G. Agraval, and A. R. Abdulov, "Phase equilibria and thermodynamics of binary copper systems with 3d-metals. vi. copper-nickel system," *Powder Metallurgy and Metal Ceramics*, vol. 46, no. 9, pp. 467–477, 2007. [Online]. Available: https://doi.org/10.1007/s11106-007-0073-x

[68] H.-C. Wang, S. Botti, and M. A. L. Marques, "Predicting stable crystalline compounds using chemical similarity," *npj Computational Materials*, vol. 7, no. 1, p. 12, 2021. [Online]. Available: https://doi.org/10.1038/s41524-020-00481-6

[69] A. Chbihi, X. Sauvage, and D. Blavette, "Atomic scale investigation of Cr precipitation in copper," *Acta Materialia*, vol. 60, no. 11, pp. 4575–4585, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359645412000766

[70] Y. Zhu, J. Liao, H. Chen, H. Wang, and B. Yang, "Solidification microstructure of Cu–Cr and Cu-Cr-In alloys," *Materials Research Express*, vol. 7, no. 4, p. 046501, apr 2020. [Online]. Available: https://dx.doi.org/10.1088/2053-1591/ab8259

[71] D. J. Chakrabarti and D. E. Laughlin, "The Cr-Cu (chromium-copper) system," *Bulletin of Alloy Phase Diagrams*, vol. 5, no. 1, pp. 59–68, 1984. [Online]. Available: https://doi.org/10.1007/BF02868727

[72] J. J. Tang, C. Liang, and C. G. Xu, "First principle calculation and thermodynamic analysis of coexisting phase of Cu-Cr-Sn copper alloy," in *Material Science and Engineering Technology X*, ser. Materials Science Forum, vol. 1053. Trans Tech Publications Ltd, 3 2022, pp. 71–76.

[73] X. Wan, W. Xie, H. Chen, F. Tian, H. Wang, and B. Yang, "First-principles study of phase transformations in Cu-Cr alloys," *Journal of Alloys and Compounds*, vol. 862, p. 158531, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838820348945

[74] H. Xie, L. Jia, and Z. Lu, "Microstructure and solidification behavior of Cu-Ni-Si alloys," *Materials Characterization*, vol. 60, no. 2, pp. 114–118, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1044580308002106

[75] S. Tao, Z. Lu, H. Xie, J. Zhang, and X. Wei, "Effect of high contents of nickel and silicon on the microstructure and properties of Cu-Ni-Si alloys," *Materials Research Express*, vol. 9, no. 4, p. 046516, apr 2022. [Online]. Available: https://dx.doi.org/10.1088/2053-1591/ac64ec

[76] J. C. Schuster and Y. Du, "Experimental investigation and thermodynamic modeling of the Cr-Ni-Si system," *Metallurgical and Materials Transactions A*, vol. 31, no. 7, pp. 1795–1803, 2000. [Online]. Available: https://doi.org/10.1007/s11661-006-0248-y

[77] X. Liu, M. Lin, S. Yang, J. Ruan, and C. Wang, "Experimental investigation of phase equilibria in the Ni-Cr-Si ternary system," *Journal of Phase Equilibria and Diffusion*, vol. 35, no. 3, pp. 334–342, 2014. [Online]. Available: https://doi.org/10.1007/s11669-014-0279-9

[78] X. Meng, G. Xie, W. Xue, Y. Fu, R. Wang, and X. Liu, "The precipitation behavior of a Cu-Ni-Si alloy with Cr addition prepared by heating-cooling combined mold (hccm) continuous casting," *Materials*, vol. 15, no. 13, 2022. [Online]. Available: https://www.mdpi.com/1996-1944/15/13/4521

[79] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," *Computational Materials Science*, vol. 58, pp. 227–235, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025612000687

[80] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013. [Online]. Available: https://doi.org/10.1063/1.4812323

[81] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design," *Acta Crystallographica Section B*, vol. 58, no. 3 Part 1, pp. 364–369, Jun 2002. [Online]. Available: https://doi.org/10.1107/S0108768102006948

[82] J. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *Journal of Metals*, vol. 65, no. 11, pp. 1501–1509, Nov. 2013.

[83] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: Review of

state of the art," *ACS Combinatorial Science*, vol. 13, no. 6, pp. 579–633, 11 2011. [Online]. Available: https://doi.org/10.1021/co200007w

[84] J. M. Sanchez, *The Cluster Expansion Method*. Boston, MA: Springer US, 1996, pp. 175–185. [Online]. Available: https://doi.org/10.1007/978-1-4613-0419-7_11

[85] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, "Data mined ionic substitutions for the discovery of new compounds," *Inorganic Chemistry*, vol. 50, no. 2, pp. 656–663, 01 2011. [Online]. Available: https://doi.org/10.1021/ic102031h

[86] D. P. Kovács, C. v. d. Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner, and G. Csányi, "Linear atomic cluster expansion force fields for organic molecules: Beyond RMSE," *Journal of Chemical Theory and Computation*, vol. 17, no. 12, pp. 7696–7711, 12 2021. [Online]. Available: https://doi.org/10.1021/acs.jctc.1c00647

[87] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, p. 146401, Apr 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.98.146401

[88] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *The Journal of Chemical Physics*, vol. 145, no. 17, p. 170901, 2023/02/01 2016. [Online]. Available: https://doi.org/10.1063/1.4966192

[89] J. Behler and G. Csányi, "Machine learning potentials for extended systems: a perspective," *The European Physical Journal B*, vol. 94, no. 7, p. 142, 2021. [Online]. Available: https://doi.org/10.1140/epjb/s10051-021-00156-1

[90] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine learning a general-purpose interatomic potential for silicon," *Phys. Rev. X*, vol. 8, p. 041048, Dec 2018. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.8.041048

[91] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, "Performance and cost assessment of machine learning interatomic potentials." *J Phys Chem A*, vol. 124, no. 4, pp. 731–745, Jan 2020.

[92] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. W. Hart, "Machine-learned multi-system surrogate models for materials prediction," *npj Computational Materials*, vol. 5, no. 1, p. 51, 2019. [Online]. Available: https://doi.org/10.1038/s41524-019-0189-9

[93] M. F. Langer, A. Goeßmann, and M. Rupp, "Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning," *npj Computational Materials*, vol. 8, no. 1, p. 41, 2022. [Online]. Available: https://doi.org/10.1038/s41524-022-00721-x

[94] A. P. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1051–1057, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24927

[95] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016. [Online]. Available: https://doi.org/10.1137/15M1054183

[96] A. Meller, L. Nivon, and D. Branton, "Voltage-driven DNA translocations through a nanopore," *Phys. Rev. Lett.*, vol. 86, no. 15, p. 3435, 2001.

[97] M. Wanunu, T. Dadosh, V. Ray, J. Jin, L. McReynolds, and M. Drndić, "Rapid electronic detection of probe-specific microRNAs using thin nanopore sensors," *Nat. Nanotech.*, vol. 5, no. 11, p. 807, 2010.

[98] A. Meller, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton, "Rapid nanopore discrimination between single polynucleotide molecules," *Proc. Nat. Ac. Sci.*, vol. 97, no. 3, pp. 1079–1084, 2000.

[99] J. Li, M. Gershow, D. Stein, E. Brandin, and J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope," *Nat. Mater.*, vol. 2, no. 9, p. 611, 2003.

[100] S. Benner, R. J. Chen, N. A. Wilson, R. Abu-Shumays, N. Hurt, K. R. Lieberman, D. W. Deamer, W. B. Dunbar, and M. Akeson, "Sequence-specific detection of individual DNA polymerase complexes in real time using a nanopore," *Nature Nanotech.*, vol. 2, no. 11, p. 718, 2007.

[101] F. J. Rang, W. P. Kloosterman, and J. de Ridder, "From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy," *Genome biology*, vol. 19, no. 1, p. 90, 2018.

[102] M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, and P. A. Pevzner, "Single-molecule protein identification by sub-nanopore sensors," *PLOS Comput. Biol.*, vol. 13, no. 5, pp. 1–14, 05 2017. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005356

[103] Q. Zhou and J. S. Liu, "Extracting sequence features to predict protein–DNA interactions: a comparative study," *Nucleic Acids Res.*, vol. 36, no. 12, pp. 4137–4148, 07 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2475627/

[104] P. Krstić, B. Ashcroft, and S. Lindsay, "Physical model for recognition tunneling," *Nanotech.*, vol. 26, no. 8, p. 084001, 2015. [Online]. Available: http://stacks.iop.org/0957-4484/26/i=8/a=084001

[105] M. Stoiber and J. Brown, "Basecrawller: Streaming nanopore basecalling directly from raw signal," *bioRxiv*, p. 133058, 2017. [Online]. Available: https://www.biorxiv.org/content/early/2017/05/01/133058

[106] M. David, L. J. Dursi, D. Yao, P. C. Boutros, and J. T. Simpson, "Nanocall: an open source basecaller for oxford nanopore sequencing data," *Bioinformatics*, vol. 33, no. 1, pp. 49–55, 2016.

[107] M. Landry and S. Winters-Hilt, "Analysis of nanopore detector measurements using machine-learning methods, with application to single-molecule kinetic analysis," *BMC Bioinformatics*, vol. 8, no. 7, p. S12, Nov 2007. [Online]. Available: https://doi.org/10.1186/1471-2105-8-S7-S12

[108] A. Churbanov and S. Winters-Hilt, "Clustering ionic flow blockade toggles with a mixture of HMMs," *BMC Bioinformatics*, vol. 9, no. 9, p. S13, Aug 2008.

[109] J. Schreiber and K. Karplus, "Analysis of nanopore data using hidden markov models," *Bioinformatics*, vol. 31, no. 12, pp. 1897–1903, 2015. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv046

[110] S. Winters-Hilt and C. Baribault, "A novel, fast, HMM-with-duration implementation - for application with a new, pattern recognition informed, nanopore detector," *BMC Bioinformatics*, vol. 8 Suppl 7, no. 7, pp. S19–S19, 11 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18047718

[111] H. Teng, M. D. Cao, M. B. Hall, T. Duarte, S. Wang, and L. J. M. Coin, "Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning," *GigaScience*, vol. 7, no. 5, pp. 1–10, 2018. [Online]. Available: http://dx.doi.org/10.1093/gigascience/giy037

[112] K. Misiunas, N. Ermann, and U. F. Keyser, "QuipuNet: Convolutional neural network for single-molecule nanopore sensing," *Nano Lett.*, vol. 18, no. 6, pp. 4040–4045, 2018. [Online]. Available: https://doi.org/10.1021/acs.nanolett.8b01709

[113] V. Boža, B. Brejová, and T. Vinař, "DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads," *PloS One*, vol. 12, no. 6, p. e0178751, 2017.

[114] R. Luo, F. J. Sedlazeck, T.-W. Lam, and M. Schatz, "Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing," *bioRxiv*, p. 310458, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/04/28/310458

[115] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, no. 1, p. 239, 2016.

[116] M.-A. Madoui, S. Engelen, C. Cruaud, C. Belser, L. Bertrand, A. Alberti, A. Lemainque, P. Wincker, and J.-M. Aury, "Genome assembly using nanopore-guided long and error-free DNA reads," *BMC Genomics*, vol. 16, no. 1, p. 327, Apr 2015. [Online]. Available: https://doi.org/10.1186/s12864-015-1519-z

[117] Y. Li, R. Han, C. Bi, M. Li, S. Wang, and X. Gao, "DeepSimulator: a deep simulator for nanopore sequencing," *Bioinformatics*, vol. 1, p. 10, 2018.

[118] P. M. Ashton, S. Nair, T. Dallman, S. Rubino, W. Rabsch, S. Mwaigwisya, J. Wain, and J. O'grady, "MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island," *Nat. Biotech.*, vol. 33, no. 3, p. 296, 2015.

[119] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," *J. Chem. Phys.*, vol. 148, no. 24, p. 241718, 2018. [Online]. Available: https://doi.org/10.1063/1.5020441

[120] C. Raillon, P. Granjon, M. Graf, L. J. Steinbock, and A. Radenovic, "Fast and automatic processing of multi-level events in nanopore translocation experiments," *Nanoscale*, vol. 4, pp. 4916–4924, 2012. [Online]. Available: http://dx.doi.org/10.1039/C2NR30951C

[121] Y. Zhang, L. Liu, J. Sha, Z. Ni, H. Yi, and Y. Chen, "Nanopore detection of DNA molecules in magnesium chloride solutions." *Nanoscale Res. Lett.*, vol. 8, p. 245, 2013.

[122] Á. Díaz Carral, C. S. Sarap, K. Liu, A. Radenovic, and M. Fyta, "2D MoS_2 nanopores: ionic current blockade height for clustering DNA events," *2D Materials*, 2019.

[123] C. Marchet, L. Lecompte, C. D. Silva, C. Cruaud, J.-M. Aury, J. Nicolas, and P. Peterlongo, "De novo clustering of long reads by gene from transcriptomics data," *Nucleic acids research*, vol. 47, no. 1, pp. e2–e2, 2018.

[124] N. Ghazikhanian, "The Matthias rules : Origins and influence," 2017.

[125] K. Conder, "A second life of the Matthias's rules," *Superconductor Science and Technology*, vol. 29, no. 8, p. 080502, jun 2016. [Online]. Available: https://dx.doi.org/10.1088/0953-2048/29/8/080502

[126] S. Chakravarty, "Quantum oscillations and key theoretical issues in high temperature superconductors from the perspective of density waves," *Reports on Progress in Physics*, vol. 74, no. 2, p. 022501, jan 2011. [Online]. Available: https://dx.doi.org/10.1088/0034-4885/74/2/022501

[127] J. Paglione and R. L. Greene, "High-temperature superconductivity in iron-based materials," *Nature Physics*, vol. 6, no. 9, pp. 645–658, 2010. [Online]. Available: https://doi.org/10.1038/nphys1759

[128] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025618304877

[129] Y. Dan, R. Dong, Z. Cao, X. Li, C. Niu, S. Li, and J. Hu, "Computational prediction of critical temperatures of superconductors based on convolutional gradient boosting decision trees," *IEEE Access*, vol. 8, pp. 57 868–57 878, 2020.

[130] B. Roter and S. V. Dordevic, "Predicting new superconductors and their critical temperatures using machine learning," *Physica C: Superconductivity and its Applications*, vol. 575, p. 1353689, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921453420301374

[131] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Phys. Rev. Lett.*, vol. 114, p. 105503, Mar 2015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.114.105503

[132] KMeans, "Kmeans," https://scikit-learn.org/stable/modules/clustering.html#k-means, Accessed: 2020-02-17.

[133] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[134] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[135] B. Roter, N. Ninkovic, and S. V. Dordevic, "Clustering superconductors using unsupervised machine learning," *Physica C: Superconductivity and its Applications*, vol. 598, p. 1354078, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921453422000661

[136] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, p. 16028, 2016. [Online]. Available: https://doi.org/10.1038/npjcompumats.2016.28

[137] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, "Machine learning modeling of superconducting critical temperature," *npj Computational Materials*, vol. 4, no. 1, p. 29, 2018. [Online]. Available: https://doi.org/10.1038/s41524-018-0085-8

[138] K. Matsumoto and T. Horide, "An acceleration search method of higher T_c superconductors by a machine learning algorithm," *Applied Physics Express*, vol. 12, no. 7, p. 073003, 2019. [Online]. Available: https://dx.doi.org/10.7567/1882-0786/ab2922

[139] Z.-L. Liu, P. Kang, Y. Zhu, L. Liu, and H. Guo, "Material informatics for layered high-T_c superconductors," *APL Materials*, vol. 8, no. 6, 06 2020, 061104. [Online]. Available: https://doi.org/10.1063/5.0004641

[140] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, and A. Maeda, "Deep learning model for finding new superconductors," *Phys. Rev. B*, vol. 103, p. 014509, Jan 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.103.014509

[141] S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, and J. Ni, "Atom table convolutional neural networks for an accurate prediction of compounds properties," *npj Computational Materials*, vol. 5, no. 1, p. 84, 2019. [Online]. Available: https://doi.org/10.1038/s41524-019-0223-y

[142] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B*, vol. 87, p. 184115, May 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.87.184115

[143] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," *Computer Physics Communications*, vol. 247, p. 106949, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010465519303042

[144] J. Zhang, Z. Zhu, X. D. Xiang, K. Zhang, S. Huang, C. Zhong, H.-J. Qiu, K. Hu, and X. Lin, "Machine learning prediction of superconducting critical temperature through the structural descriptor," *The Journal of Physical Chemistry C*, vol. 126, no. 20, pp. 8922–8927, 05 2022. [Online]. Available: https://doi.org/10.1021/acs.jpcc.2c01904

[145] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun *et al.*, "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *Lancet*, vol. 8, no. 4, pp. e488 – e496, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214109X20300747

[146] J. Bedford, D. Enria, J. Giesecke, D. L. Heymann, C. Ihekweazu, G. Kobinger, H. C. Lane, Z. Memish, M. don Oh, A. A. Sall *et al.*, "COVID-19: towards controlling of a pandemic," *Lancet*, vol. 395, no. 10229, pp. 1015 – 1018, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673620306735

[147] J. Hopman, B. Allegranzi, and S. Mehtar, "Managing COVID-19 in Low- and Middle-Income Countries," *JAMA, J. Am. Med. Assoc.*, vol. 323, no. 16, pp. 1549–1550, 04 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.4169

[148] M. I. Yatoo, Z. Hamid, O. R. Parray, A. H. Wani, A. U. Haq, A. Saxena, S. K. Patel, M. Pathak, R. Tiwari, Y. S. Malik *et al.*, "COVID-19 - recent advancements in identifying novel vaccine candidates and current status of upcoming SARS-CoV-2 vaccines," *Hum. Vaccines Immunother.*, vol. 0, no. 0, pp. 1–14, 2020, pMID: 32703064. [Online]. Available: https://doi.org/10.1080/21645515.2020.1788310

[149] J. R. Brister, D. Ako-adjei, Y. Bao, and O. Blinkova, "NCBI viral genomes resource," *Nucleic Acids Res.*, vol. 43, p. D571–D577, 2015.

[150] L. Dong, S. Hu, and J. Gao, "Discovering drugs to treat coronavirus disease 2019 (COVID-19)," *Drug Discoveries Ther.*, vol. 14, no. 1, pp. 58–60, 2020.

[151] B. N. Rome and J. Avorn, "Drug evaluation during the COVID-19 pandemic," *N. Engl. J. Med.*, vol. 382, no. 24, pp. 2282–2284, 2020. [Online]. Available: https://doi.org/10.1056/NEJMp2009457

[152] B. Udugama, P. Kadhiresan, H. N. Kozlowski, A. Malekjahani, M. Osborne, V. Y. C. Li, H. Chen, S. Mubareka, J. B. Gubbay, and W. C. W. Chan, "Diagnosing COVID-19: The disease and tools for detection," *ACS Nano*, vol. 14, no. 4, pp. 3822–3835, 2020, pMID: 32223179. [Online]. Available: https://doi.org/10.1021/acsnano.0c02624

[153] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, D. G. J. C. Mulders, R. Molenkamp, C. A. Perez-Romero, E. Claassen, J. Garssen, and A. D. Kraneveld, "Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning," *Sci. Rep.*, vol. 11, no. 1, p. 947, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-80363-5

[154] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[155] Y. Hozumi, R. Wang, C. Yin, and G.-W. Wei, "UMAP-assisted k-means clustering of large-scale SARS-CoV-2 mutation datasets," *Computers in biology and medicine*, vol. 131, p. 104264, 2021.

[156] G. D'Angelo and F. Palmieri, "Discovering genomic patterns in SARS-CoV-2 variants," *Int. J. Intell. Syst.*, vol. 35, no. 11, pp. 1680–1698, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22268

[157] T. A. Brown, *Genomes*, 2nd ed. Oxford: Wiley-Liss, 2002.

[158] P. Sieber, M. Platzer, and S. Schuster, "The definition of open reading frame revisited," *Trends Genet.*, vol. 34, no. 3, pp. 167 – 170, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168952517302299

[159] C. W. Nelson, Z. Ardern, T. L. Goldberg, C. Meng, C.-H. Kuo, C. Ludwig, S.-O. Kolokotronis, and X. Wei, "Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic," *eLife*, vol. 9, p. e59633, 2020.

[160] J. Z. Tan Yao, "Visualizations of multiple probability measures for SARS-CoV-2 genomes," vol. PREPRINT available at Research Square, pp. https://doi.org/10.21 203/rs.3.rs–74 631/v1+, 09 2020.

[161] Y. Wang, K. Tian, and S. S.-T. Yau, "Protein sequence classification using natural vector and convex hull method," *Journal of Computational Biology*, vol. 26, no. 4, pp. 315–321, 2019.

[162] Y. Van Der Meer, H. van Tol, J. K. Locker, and E. J. Snijder, "ORF1a-encoded replicase subunits are involved in the membrane association of the arterivirus replication complex," *J. Virol.*, vol. 72, no. 8, pp. 6689–6698, 1998.

[163] E. Méndez, M. E. Salas-Ocampo, M. E. Munguía, and C. F. Arias, "Protein products of the open reading frames encoding nonstructural proteins of human astrovirus serotype 8," *J. Virol.*, vol. 77, no. 21, pp. 11 378–11 384, 2003.

[164] R. L. Graham, J. S. Sparks, L. D. Eckerle, A. C. Sims, and M. R. Denison, "SARS coronavirus replicase proteins in pathogenesis," *Virus Res.*, vol. 133, no. 1, pp. 88 – 100, 2008, sARS-CoV Pathogenesis and Replication. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168170207000573

[165] R. Hershberg and D. Petrov, "Selection on codon bias," *Annu. Rev. Genet.*, vol. 42, pp. 287–99, 02 2008.

[166] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.

[167] P. Martins, "Gene prediction using deep learning," 2018. [Online]. Available: https://hdl.handle.net/10216/114372

[168] X. Zhu, J. Lafferty, and R. Rosenfeld, "Semi-supervised learning with graphs," Ph.D. dissertation, USA, 2005, aAI3179046.

[169] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. NLD: IOS Press, 2007, p. 3–24.

[170] D. Greene, P. Cunningham, and R. Mayer, *Unsupervised Learning and Clustering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 51–90. [Online]. Available: https://doi.org/10.1007/978-3-540-75171-7_3

[171] O. Chapelle, B. Scholkopf, and A. Zien, Eds., "Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[172] S. Becker and M. Plumbley, "Unsupervised neural network learning procedures for feature extraction and classification," *Applied Intelligence*, vol. 6, no. 3, pp. 185–203, Jul. 1996. [Online]. Available: https://doi.org/10.1007/BF00126625

[173] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Int. Res.*, vol. 4, no. 1, p. 237–285, may 1996.

[174] D. Cohn, "Active learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer, 2011.

[175] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. USA: Prentice-Hall, Inc., 1988.

[176] H. J. Miller and J. Han, Eds., *Geographic Data Mining and Knowledge Discovery*, 1st ed. CRC Press, 2001.

[177] J. J. Shen, P.-H. Lee, J. J. A. Holden, and H. Shatkay, "Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders," in *AMIA Annual Symposium Proceedings*, vol. 2007, 2007, pp. 666–670. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18693920

[178] T. Xięski, A. Nowak-Brzezińska, and A. Wakulicz-Deja, "Density-based method for clustering and visualization of complex data," in *Rough Sets and Current Trends in Computing*, J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra, and L. Polkowski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 142–149.

[179] M. R. Ilango and V. Mohan, "A survey of grid-based clustering algorithms," *International Journal of Engineering Science and Technology*, vol. 2, 2010.

[180] A. Zhang, "Acceleration of k-means clustering by k-dijkstra method for graph partitioning (master thesis)," Master's thesis, 2015, aAI3179046.

[181] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, "The clustering validity with silhouette and sum of squared errors (proceedings)," 2015, pp. 44–51.

[182] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[183] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Wadsworth, 1984.

[184] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986. [Online]. Available: https://doi.org/10.1007/BF00116251

[185] B. F. Murorunkwere, J. F. Ihirwe, I. Kayijuka, J. Nzabanita, and D. Haughton, "Comparison of tree-based machine learning algorithms to predict reporting behavior of electronic billing machines," *Information*, vol. 14, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/3/140

[186] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" 2022.

[187] G. Louppe, "Understanding Random Forests: From theory to practice," 2015.

[188] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[189] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 10 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[190] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021

[191] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *arXiv preprint arXiv:1603.02754*, 2016.

[192] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016, http://www.deeplearningbook.org.

[193] B. G. Farley and W. A. Clark, "Simulation of self-organizing systems by digital computer," *Transactions of the IRE Professional Group on Information Theory (TIT)*, vol. 4, pp. 76–84, 1954.

[194] C. von der Malsburg, "Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," in *Brain Theory*, 1986, pp. 245–248.

[195] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[196] N. Baker, F. Alexander, T. Bremer, A. Hagberg, Y. Kevrekidis, H. Najm, M. Parashar, A. Patra, J. Sethian, S. Wild *et al.*, "Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence." [Online]. Available: https://www.osti.gov/biblio/1478744

[197] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics–informed neural networks: Where we are and what is next," *Journal of Scientific Computing*, vol. 92, no. 3, p. 88, Jul. 2022. [Online]. Available: https://doi.org/10.1007/s10915-022-01939-z

[198] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021.

[199] S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao, and B. Blum-Smith, "Scalars are universal: Equivariant machine learning, structured like classical physics," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 28 848–28 863, 6 2021.

[200] M. Melchior, "Incorporating domain knowledge for learning interpretable features," *Archives of Data Science, Series A (Online First)*, vol. 8, no. 2, p. 14 S. online, 2022.

[201] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[202] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations," 2021.

[203] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J Physiol*, vol. 160, no. 1, pp. 106–154, Jan. 1962.

[204] S. Sonnenburg, "Machine learning for genomic sequence analysis." 01 2008, pp. 281–290.

[205] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 10 2000. [Online]. Available: https://doi.org/10.1162/089976600300015015

[206] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[207] K. Choudhary, F. Y. P. Congo, T. Liang, C. Becker, R. G. Hennig, and F. Tavazza, "Evaluation and comparison of classical interatomic potentials through a user-friendly interactive web-interface," *Scientific Data*, vol. 4, no. 1, p. 160125, Jan. 2017. [Online]. Available: https://doi.org/10.1038/sdata.2016.125

[208] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B*, vol. 87, p. 184115, May 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.87.184115

[209] C. Chen and S. P. Ong, "A universal graph deep learning interatomic potential for the periodic table," *Nature Computational Science*, vol. 2, no. 11, pp. 718–728, Nov. 2022. [Online]. Available: https://doi.org/10.1038/s43588-022-00349-3

[210] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," *Journal of Computational Physics*, vol. 285, pp. 316–330, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999114008353

[211] P. van Gerwen, A. Fabrizio, M. D. Wodrich, and C. Corminboeuf, "Physics-based representations for machine learning properties of chemical reactions," *Machine Learning: Science and Technology*, vol. 3, no. 4, p. 045005, oct 2022. [Online]. Available: https://dx.doi.org/10.1088/2632-2153/ac8f1a

[212] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "The MLIP package: moment tensor potentials with MPI and active learning," *Machine Learning: Science and Technology*, vol. 2, no. 2, p. 025002, dec 2020. [Online]. Available: https://dx.doi.org/10.1088/2632-2153/abc9fe

[213] K. Burke, *The ABC of DFT*. CA 92697: Department of Chemistry, University of California, Irvine, 2007.

[214] E. Kaxiras, *Atomic and Electronic Structure of Solids*. Cambridge University Press, 2003.

[215] P. Bouř, "Comparison of Hartree–Fock and Kohn–Sham determinants as wave functions," *Journal of Computational Chemistry*, vol. 21, pp. 8 – 16, 01 2000.

[216] J. Thijssen, *Computational Physics*, 2nd ed. Cambridge University Press, 2007.

[217] H. J. Berendsen, *Simulating the Physical World*. Cambridge University Press, 2007, iSBN: 0-521-83527-5.

[218] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. San Diego: Academic Press, 2002.

[219] L. H. Thomas, "The calculation of atomic fields," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 23, no. 5, p. 542–548, 1927.

[220] E. Fermi, "Un metodo statistico per la determinazione di alcune prioprietà dell'atomo," *Accademia Nazionale dei Lincei*, vol. 6, pp. 602–607, 1927.

[221] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, pp. B864–B871, Nov 1964. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.136.B864

[222] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, pp. A1133–A1138, Nov 1965. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.140.A1133

[223] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conference Proceedings*, vol. 577, no. 1, pp. 1–20, 07 2001. [Online]. Available: https://doi.org/10.1063/1.1390175

[224] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, Oct 1996. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865

[225] R. E. A. Goodall, "Accelerating materials discovery with machine learning," 2022. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/335048

[226] M. I. Baskes, "Modified embedded-atom potentials for cubic materials and impurities," *Physical review B*, vol. 46, no. 5, p. 2727, 1992.

[227] B. Jelinek, S. Groh, M. F. Horstemeyer, J. Houze, S.-G. Kim, G. J. Wagner, A. Moitra, and M. I. Baskes, "Modified embedded atom method potential for Al, Si, Mg, Cu, and Fe alloys," *Physical Review B*, vol. 85, no. 24, p. 245102, 2012.

[228] M. S. Daw and M. Baskes, "Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals," *Physical Review Letters*, vol. 50, no. 17, p. 1285 – 1288, 1983, cited by: 2454. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-4244024430&doi=10.1103%2fPhysRevLett.50.1285&partnerID=40&md5=83e36a726ea7a4115b1ffa32c149ad6e

[229] M. W. Finnis and J. E. Sinclair, "A simple empirical n-body potential for transition metals," *Philosophical Magazine A*, vol. 50, no. 1, pp. 45–55, 1984. [Online]. Available: https://doi.org/10.1080/01418618408244210

[230] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Computer Physics Communications*, vol. 271, p. 108171, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010465521002836

[231] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics*, vol. 76, no. 1, pp. 637–649, 01 1982. [Online]. Available: https://doi.org/10.1063/1.442716

[232] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of Chemical Physics*, vol. 81, no. 1, pp. 511–519, 07 1984. [Online]. Available: https://doi.org/10.1063/1.447334

[233] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, pp. 1695–1697, Mar 1985. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.31.1695

[234] G. J. Martyna, M. L. Klein, and M. Tuckerman, "Nosé–Hoover chains: The canonical ensemble via continuous dynamics," *The Journal of Chemical Physics*, vol. 97, no. 4, pp. 2635–2643, 08 1992. [Online]. Available: https://doi.org/10.1063/1.463940

[235] W. S. Morgan, G. L. Hart, and R. W. Forcade, "Generating derivative superstructures for systems with high configurational freedom," *Computational Materials Science*, vol. 136, pp. 144–149, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025617302069

[236] G. L. W. Hart and R. W. Forcade, "Algorithm for generating derivative structures," *Phys. Rev. B*, vol. 77, p. 224115, Jun 2008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.77.224115

[237] L. Vegard, "Die konstitution der mischkristalle und die raumfüllung der atome," *Zeitschrift für Physik*, vol. 5, no. 1, pp. 17–26, 1921. [Online]. Available: https://doi.org/10.1007/BF01349680

[238] A. R. Denton and N. W. Ashcroft, "Vegard's law," *Phys. Rev. A*, vol. 43, pp. 3161–3164, Mar 1991. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.43.3161

[239] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy *et al.*, "AFLOW: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025612000717

[240] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, p. 3865, 1996.

[241] P. E. Blöchl, "Projector augmented-wave method," *Physical review B*, vol. 50, no. 24, p. 17953, 1994.

[242] A. Togo and I. Tanaka, "First principles phonon calculations in materials science," *Scripta Materialia*, vol. 108, pp. 1–5, 2015.

[243] I. Pallikara, P. Kayastha, J. M. Skelton, and L. D. Whalley, "The physical significance of imaginary phonon modes in crystals," *Electronic Structure*, vol. 4, no. 3, p. 033002, 2022. [Online]. Available: https://dx.doi.org/10.1088/2516-1075/ac78b3

[244] X. Zhang, B. Grabowski, F. Körmann, A. V. Ruban, Y. Gong, R. C. Reed, T. Hickel, and J. Neugebauer, "Temperature dependence of the stacking-fault gibbs energy for Al, Cu, and Ni," *Phys. Rev. B*, vol. 98, p. 224106, Dec 2018. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.98.224106

[245] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," *Nature Communications*, vol. 8, no. 1, p. 15679, 2017. [Online]. Available: https://doi.org/10.1038/ncomms15679

[246] E. Gossett, C. Toher, C. Oses, O. Isayev, F. Legrain, F. Rose, E. Zurek, J. Carrete, N. Mingo, A. Tropsha *et al.*, "AFLOW-ML: A RESTful API for machine-learning predictions of materials properties," *Computational Materials Science*, vol. 152, pp. 134–145, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025618302349

[247] C.-Y. Wen and F. Spaepen, "In situ electron microscopy of the phases of Cu_3Si," *Philosophical Magazine*, vol. 87, no. 35, pp. 5581–5599, 2007. [Online]. Available: https://doi.org/10.1080/14786430701675829

[248] The Materials Project, "Materials data on Si_2Ni_3 by materials project," 7 2020.

[249] R. Gaillac, P. Pullumbi, and F.-X. Coudert, "ELATE: an open-source online application for analysis and visualization of elastic tensors," *Journal of Physics: Condensed Matter*, vol. 28, no. 27, p. 275201, may 2016. [Online]. Available: https://dx.doi.org/10.1088/0953-8984/28/27/275201

[250] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne, "First principles methods using CASTEP," *Zeitschrift für Kristallographie - Crystalline Materials*, vol. 220, no. 5-6, pp. 567–570, 2005. [Online]. Available: https://doi.org/10.1524/zkri.220.5.567.65075

[251] D. Hicks, C. Oses, E. Gossett, G. Gomez, R. H. Taylor, C. Toher, M. J. Mehl, O. Levy, and S. Curtarolo, "*AFLOW-SYM*: platform for the complete, automatic and self-consistent symmetry analysis of crystals," *Acta Crystallographica Section A*, vol. 74, no. 3, pp. 184–203, May 2018. [Online]. Available: https://doi.org/10.1107/S2053273318003066

[252] K. Dies, *Legierungen des Kupfers mit Elementen der 4. Gruppe des Periodischen Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1967, pp. 482–655. [Online]. Available: https://doi.org/10.1007/978-3-642-48931-0_8

[253] I. Novikov, B. Grabowski, F. Körmann, and A. Shapeev, "Magnetic moment tensor potentials for collinear spin-polarized materials reproduce different magnetic states of bcc fe," *npj Computational Materials*, vol. 8, no. 1, p. 13, 2022.

[254] C. J. Bartel, "Review of computational approaches to predict the thermodynamic stability of inorganic solids," *Journal of Materials Science*. [Online]. Available: https://www.osti.gov/biblio/1865532

[255] R. Olesinski and G. Abbaschian, "The Cu-Si (copper-silicon) system," *Bulletin of Alloy Phase Diagrams*, vol. 7, no. 2, pp. 170–178, 1986.

[256] B. Jelinek, S. Groh, M. F. Horstemeyer, J. Houze, S. G. Kim, G. J. Wagner, A. Moitra, and M. I. Baskes, "Modified embedded atom method potential for Al, Si, Mg, Cu, and Fe alloys," *Phys. Rev. B*, vol. 85, p. 245102, Jun 2012. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.85.245102

[257] J. Feng, K. Liu, R. D. Bulushev, S. Khlybov, D. Dumcenco, A. Kis, and A. Radenovic, "Identification of single nucleotides in MoS_2 nanopores," *Nature Nanotech.*, vol. 10, p. 1070, 09 2015. [Online]. Available: http://dx.doi.org/10.1038/nnano.2015.219

[258] M. Graf, M. Lihter, D. Altus, S. Marion, and A. Radenovic, "Transverse detection of DNA using a MoS_2 nanopore," *Nano letters*, vol. 19, no. 12, pp. 9075–9083, 2019.

[259] C. Raillon, P. Granjon, M. Graf, L. Steinbock, and A. Radenovic, "Fast and automatic processing of multi-level events in nanopore translocation experiments," *Nanoscale*, vol. 4, pp. 4916–24, 07 2012.

[260] J. M. Schreiber and K. Karplus, "Segmentation of noisy signals generated by a nanopore," *bioRxiv*, p. 014258, 2015. [Online]. Available: https://www.biorxiv.org/content/early/2015/01/23/014258

[261] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[262] C. Cengizler and M. Kerem Un, "Evaluation of Calinski-Harabasz criterion as fitness measure for genetic algorithm based segmentation of cervical cell nuclei," *British J. Math. Comput. Sci.*, vol. 22, no. 6, pp. 1–13, 2017.

[263] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," *IEEE ICDM*, pp. 911–916, 2010.

[264] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[265] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, July 2002.

[266] P. C. Faucon, R. Trevino, P. Balachandran, K. Standage-Beier, and X. Wang, "High accuracy base calls in nanopore sequencing," *bioRxiv*, p. 126680, 2017. [Online]. Available: https://www.biorxiv.org/content/early/2017/04/11/126680

[267] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, 11 2000.

[268] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[269] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[270] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.4.541

[271] L. A. Santamaría, S. Zuñiga, I. H. Pineda, M. J. Somodevilla, and M. Rossainz, "Reconocimiento de genes en secuencias de ADN por medio de imágenes," in *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018): avances en Inteligencia Artificial*, F. Herrera Triguero, A. Troncoso Lara, and S. Damas Arroyo, Eds., Granada, España, octubre 2018, pp. 808–812.

[272] B. Antonescu, M. Tehrani Moayyed, and S. Basagni, "Clustering algorithms and validation indices for a wide mmwave spectrum," *Information*, vol. 10, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2078-2489/10/9/287

[273] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on artificial intelligence*, vol. 56, 2000, pp. 111–117.

[274] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97, no. 1. Citeseer, 1997, p. 179.

[275] Á. Díaz Carral, M. Roitegui, A. Koc, M. Ostertag, and M. Fyta, "Concurrent analysis of electronic and ionic nanopore signals: blockade mean and height," *Nano Express*, vol. 5, no. 2, p. 025020, jun 2024. [Online]. Available: https://dx.doi.org/10.1088/2632-959X/ad4dbf

[276] "GitHub - zhanzghang/superconductors-data — github.com," https://github.com/zhanzghang/superconductors-data, [Accessed 05-12-2023].

[277] M. Pop, G. Borodi, and S. Simon, "Correlation between valence electron concentration and high-temperature superconductivity," *Journal of Physics and Chemistry of Solids*, vol. 61, no. 12, pp. 1939–1944, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022369700000846

[278] A. Liaw and M. Wiener, "Classification and regression by Random Forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[279] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*. Cambridge, MA, USA: MIT Press, 1986, p. 318–362.

[280] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[281] S. Lundberg, "SHAP," 2018, accessed on Jun 27, 2023. [Online]. Available: https://github.com/slundberg/shap

[282] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9

[283] B. T. Matthias, T. H. Geballe, R. H. Willens, E. Corenzwit, and G. W. Hull, "Superconductivity of Nb_3Ge," *Phys. Rev.*, vol. 139, pp. A1501–A1503, Aug 1965. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.139.A1501

[284] Á. Díaz Carral, M. Roitegui, and M. Fyta, "Interpretably learning the critical temperature of superconductors: Electron concentration and feature dimensionality reduction," *APL Materials*, vol. 12, no. 4, p. 041111, 6/12/2024 2024. [Online]. Available: https://doi.org/10.1063/5.0189714

[285] National Center for Biotechnology Information (NCBI), "National center for biotechnology information," accessed: 2020-10-13. [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/viruses/

[286] J. P. Sarkar, I. Saha, A. Seal, and D. Maity, "COVID-Predictor: RNA sequence based prediction of coronavirus," vol. PREPRINT (Version 1) available at Research Square, pp. https://doi.org/10.21 203/rs.3.rs–23 913/v1, 2020. [Online]. Available: https://doi.org/10.21203/rs.3.rs-23913/v1

[287] CovidPredictor, "COVID-Predictor: Machine learning to predict novel coronavirus from other pathogenic viruses," accessed: 2020-06-16. [Online]. Available: http://www.nitttrkol.ac.in/indrajit/projects/COVID-Predictor/

[288] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009. [Online]. Available: https://doi.org/10.1093/bioinformatics/btp163

[289] J.-M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences," *Hum. Mol. Genet.*, vol. 6, no. 10, pp. 1735–1744, 09 1997. [Online]. Available: https://doi.org/10.1093/hmg/6.10.1735

[290] I. T. Rombel, K. F. Sykes, S. Rayner, and S. A. Johnston, "ORF-FINDER: a vector for high-throughput gene identification," *Gene*, vol. 282, no. 1, pp. 33 – 41, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378111901008198

[291] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith, "Cython: The best of both worlds," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 31–39, 2011.

[292] M. Hahsler, M. Piekenbrock, and D. Doran, "DBSCAN: Fast density-based clustering with R," *J. Stat. Softw*, vol. 91, no. 1, pp. 1–30, 2019.

[293] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–136, 1982.

[294] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE PAMI*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[295] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," *IEEE International Conference on Data Mining, ICDM*, pp. 187–194, 02 2001.

[296] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," *LNCS (LNAI)*, vol. 1910, pp. 265–276, 01 2000.

[297] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," *IEEE ICDM*, 2011.

[298] L. Oberer, Á. Díaz Carral, and M. Fyta, "Simple classification of RNA sequences of respiratory-related coronaviruses," *ACS Omega*, vol. 6, no. 31, pp. 20 158–20 165, Aug. 2021, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acsomega.1c01625