UNIVERSITY OF BLIDA 1

FACULTY OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

DISSERTATION

Option: Computer and Data Science

Crystal Structure Prediction Using Data Mining Techniques/Through a Neural Approach

---

Presented by Mouzai Meriem

Before the jury:

N. Benblidia	**Professor,**	**U. of Blida 1**	**President**
S. Oukid	**Professor,**	**U. of Blida 1**	**Supervisor**
A. Oganov	**Professor,**	**Skoltech, Moscow**	**Co-Supervisor**
M. Fareh	**Lecturer Class A,**	**U. of Blida 1**	**Examiner**
E. Garoudja	**Sr. Researcher Class A,**	**CDTA, Algiers**	**Examiner**

June 2025

To my beloved parents

Acknowledgements

Praise be to the Almighty God who has given me faith, courage, and patience to carry out this work.

I would like to express my deep gratitude to my supervisor Prof. Saliha Oukid whose constant support and direction were essential throughout this thesis adventure. Her encouragement to consider many options and avenues as well as her knowledgeable advice were really helpful in achieving our study objectives. Her commitment and insight, which were crucial in realizing this thesis, are greatly appreciated.

I would like to extend my sincere appreciation to my co-supervisor Prof. Artem Oganov. His generosity in welcoming me into his research lab and his provision of the priceless resources I required has been fundamental in making this thesis a reality. I am deeply grateful for his collaborative spirit, his guidance, unwavering availability to address my questions, and willingness to share his expertise which have significantly enriched my research journey.

I would also like to extend my sincere thanks to the members of the jury for kindly agreeing to evaluate my thesis. I am honored by their involvement and grateful for the time and attention they dedicated to the assessment of my work.

I would like to express my gratitude to all those who assisted me in enhancing my work, particularly my colleagues from CDTA, including Hania Djani for her priceless support and guidance during the course of this research, and the doctoral school, whose feedback and encouragement contributed in one way or another to the development of this thesis. I am also deeply thankful to Alaeddine Oulahcene for his unwavering moral support and encouragement, especially during the final stretch of preparing my defense.

ملخص

يعدّ التنبؤ بالطاقة في البنيات البلورية باستخدام الذكاء الاصطناعي من مجالات البحث المهمة لميداني علوم المواد والصناعة. تقدّم هذه الأطروحة دراسة متعددة التخصصات من أجل تعزيز المفاهيم العلمية وتطبيقاتها في الحياة اليومية، مما يعزز الابتكار ويمكّن الباحثين والمصنعين على السعي لتطوير تقنيات أكثر تقدماً واستدامة. توفر هذه الدراسة أدوات دقيقة لتصميم المواد بخصائص محددة، مما يقلل إلى حد كبير من مدة الإنجاز مقارنةً بالأساليب التقليدية. تهدف هذه الدراسة إلى استخدام نماذج الذكاء الاصطناعي (تعلم الآلة والتعلم العميق) كبديل فعّال للطريقة المختبرية. لتحقيق هذه الغاية، تمّ تحويل البيانات الأولية المعقدة والخام للمواد المجمعة إلى مدخلات قابلة للقراءة من قبل الآلة، من خلال استخدام دوال التوزيع الثنائية والثلاثية للجسم الذري، لاستخراج الخصائص البنيوية والذرية للبيانات البلورية المجمعة من أجل تحويلها إلى مدخلات يمكن قراءتها آلياً. بعد ذلك، تمّ استخدام خوارزميات الذكاء الاصطناعي، التي تشمل الشبكة العصبية العميقة، آلة الدعم الناقل، الغابة العشوائية، والانحدار البايزي والمرن، لتمثيل العلاقة بين خاصية الطاقة والمدخلات البنيوية. علاوة على ذلك، تمّ اقتراح وتنفيذ بنية غير تقليدية للشبكة العصبية العميقة لدعم المميزات الذرية. في مرحلة لاحقة، تمّ ضبط المعاملات الأساسية للنماذج المقترحة من أجل الحصول على أحسن أداء ممكن. بالإضافة إلى ذلك، تمّ تطبيق مقاييس التقييم من أجل اختبار النماذج واختيار المميز الأقوى للنموذج المتحصل عليه في التنبؤ بالطاقة. أظهرت النتائج المتحصلة علمياً من هذه الأطروحة أن الجمع بين دوال التوزيع الذرية الثنائية والثلاثية للجسم، ونموذج الشبكة العصبية المقترح، أعطى أدق وأفضل النتائج.

كلمات مفتاحية : الذكاء الاصطناعي، تعلم الآلة، التعلم العميق، مزايا البنيات البلورية، التنبؤ بالطاقة.

Abstract

Crystal structure energy prediction with Artificial Intelligence (AI) algorithms is a significant research for both materials science and industry. This thesis reports a multidisciplinary study to enhance scientific understanding and real-world applications, making it important for researchers and industries seeking more effective technologies. It provides accurate tools for designing materials with tailored properties, considerably reducing the time consuming and resource-intensive testing of conventional approaches. This work investigates artificial intelligence models (machine learning-ML and deep learning-DL) to substitute the laboratory crystal structure energy prediction. To this end, two- and three-body distribution functions were used to transform raw, complex material details of the collected data into machine-readable inputs, resulting in structural and atomic descriptors. Then, ML/DL algorithms, namely: ElasticNet, Bayesian ridge, random forest, support vector machine, and deep neural networks were used to model the relationship between the energy property and the structural descriptors. Moreover, a non-conventional deep neural networks topology was proposed and implemented to support atomic descriptors. Hyper-parameter tuning was performed on each model for optimization purpose. Additionally, quality assessment metrics were used to test and evaluate the energy prediction yielded by the investigated models in order to select the most robust descriptors and the best performing model. The obtained results revealed that the most accurate energy prediction was achieved by combining two- and three body atomic distribution functions as a descriptor, and the proposed deep neural networks model.

***Key words**: Artificial Intelligence, Machine Learning, Deep Learning, Crystal Structure Features Descriptors, Energy Prediction*.

Résumé

La prédiction de la propriété d'énergie des structures cristallines en utilisant l'intelligence artificielle est un domaine de recherche significatif, à la fois pour la science des matériaux et l'industrie. Cette étude multidisciplinaire améliore la compréhension scientifique et présente des applications réelles, ce qui la rend importante pour les chercheurs et les industries à la recherche de technologies plus efficaces. Elle fournit des outils précis pour concevoir des matériaux aux propriétés sur mesure, réduisant considérablement le temps d'exécution et les ressources coûteuses nécessaires aux tests traditionnels. Ce travail explore des modèles d'intelligence artificielle (apprentissage automatique et apprentissage profond) pour remplacer la prédiction de l'énergie des structures cristallines en laboratoire. A cette fin, des fonctions de distribution à deux et trois corps ont été utilisées pour transformer les détails complexes et bruts des matériaux collectés en entrées lisibles par machine, résultant en des descripteurs structurels et atomiques. Ensuite, des algorithmes d'apprentis sage automatique/profond, à savoir ElasticNet, Bayesian ridge, forêt d'arbre de décision, machine à vecteurs de support et réseaux de neurones profonds, ont été utilisés pour modéliser la relation entre la propriété de l'énergie et les descripteurs structurels. De plus, une topologie non conventionnelle de réseaux de neurones profonds a été proposée et implémentée pour prendre en charge les descripteurs atomiques. Un ajustement des hyperparamètres a été réalisé sur chaque modèle à des fins d'optimisation. De plus, des métriques d'évaluation ont été utilisées pour tester et évaluer la prédiction d'énergie obtenue par les modèles étudiés afin d'identifier les descripteurs les plus robustes et le modèle le plus performant. Les résultats de cette étude ont révélé que la prédiction d'énergie la plus précise a été obtenue en utilisant la combinaison des fonctions de distribution atomique à deux et trois corps en tant que descripteur, ainsi que le modèle de réseaux de neurones profonds proposé.

Mots clés *: Intelligence Artificielle, Apprentissage Automatique, Apprentissage Profond, Descripteurs de Caractéristiques des Structures Cristallines, Prédiction d'énergie.*

Contents

List of Figures

Note: All figures in this thesis were created by the author unless otherwise stated.

List of Tables

List of Abbreviations

AFS: Angular Fourier Series
AI: Artificial Intelligence
ANN: Artificial Neural Network
BOP: Bond-orientational Order Parameter
BO: Bayesian Optimization
CSP: Crystal Structure Prediction
COD: Crystallography Open Database
DFT: Density Functional Theory
DL: Deep Learning
DOS: Density Of electronic States
FRAM: Ferroelectric Random Access Memory
GGA: Generalized Gradient Approximation
GRDF: Generalized Radial Distribution Function
GS: Ground State
HF: Hartree-Fock
ICSD: Inorganic Crystal Structure Database
KS-DFT: Kohn-Sham Density Functional Theory
KRR: Kernel Ridge Regression
LD: Local Density
LDA: Local Density Approximation
LSDA: Local Spin Density Approximation
MEM: Microelectromechanical
ML: Machine Learning
MP: Materials Project
PRDF: Partial Radial Distribution Function
TFD: Thomas-Fermi-Dirac
XRD: X-ray Diffraction

Adam:	ADAptive Moment estimation
AFP:	Atomic Fingerprints
AFLOW:	Automatic FLOW for materials discovery
AUC:	Area Under (ROC) Curve
BFGS:	Broyden-Fletcher-Goldfarb-Shanno
CGCNN:	Crystal Graph Convolutional Neural Network
CNN:	Convolutional Neural Networks
CSFP:	Crystal Structure FingerPrint
DNN:	Deep Neural Networks
ECOC:	Error-Correcting Output Coding
ExRT:	EXtremely Randomized Trees
GA:	Genetic Algorithm
GBDT:	Gradient Boosting Decision Tree
GBM:	Gradient Boosting Machine
GBRT:	Gradient Boosted Regression Trees
GAT:	Graph Attention
GNN:	Graph Neural Networks
HEA :	High Entropy Alloys
KISS:	Keep It Simple and Straightforward
KKN :	K-Nearest Neighbor
KKR-CPA:	Korringa-Kohn-Rostoker-Coherent Potential Approximation
LR :	Logistic Regression
LTC:	Lattice Thermal Conductivity
Magpie:	Materials-AGnostic Platform for Informatics and Exploration
MLIP:	Machine Learning Interatomic Potentials
MLP:	Multi-Layer Perceptron
NB:	Naïve Bayes
OQMD:	Open Quantum Materials Database
RBF:	Radial Basis Function
ReLU:	REctified Linear Unit
RF:	Random Forest
RMSD:	Root Mean Square Distance
SACADA:	SAmara Carbon Allotrope DAtabase

SGD: Stochastic Gradient Descent
SVR: Support Vector Regressor
SVM: Support Vector Machine
TF: Transfer Function
USPEX: Universal Structure Predictor: Evolutionary Xtallography
VAE: Variational AutoEncoder
VASP: Vienna Ab initio Simulation Package
XGBoost: eXtreme Gradient BOOSTing
2BDF: Two-Body Distribution Function
3BDF: Three-Body Distribution Function
2-3BDF: Two- and Three-Body Distribution Functions combined
2BDF-St: Structural Two-Body Distribution Function
3BDF-St: Structural Three-Body Distribution Function
2-3BDF-St: Structural Two- and Three-Body Distribution Functions combined
2BDF-At: Atomic Two-Body Distribution Function
3BDF-At: Atomic Three-Body Distribution Function
2-3BDF-At: Atomic Two- and Three-Body Distribution Functions combined
AGAT: Augmented Graph Attention
CIF: Crystallographic Information File
EN: ElasticNet
GNN: Graph Neural Networks
LA: Learning Algorithm
LASSO: Least Absolute Shrinkage and Selection Operator
MAE: Mean Absolute Error
MatB: MatBench
MSE: Mean Squared Error
MTP: Moment Tensor Potential
NLP: Natural Language Processing
PCA: Principal Component Analysis
PES: Potential Energy Surface
PND: Powder Neutron Diffraction
PSO: Particle Swarm Optimization
RAS: Random Searching
RMSE: Root Mean Square Error

LF: Loss Function
5F-CV: Five-Fold Cross-Validation
ROC: Receiver Operating Characteristics
TP: True Positive
FN: False Negative
TN: True Negative
FP: False Positive
TPR: True Positive Rate
FPR: False Positive Rate

General Introduction

Context and motivation

In the realm of materials science, an intricate landscape made up of threads of atoms and molecules, crystal structures are the masterpieces unlocking remarkable properties and applications. One of the biggest ambitions of materials scientists and researchers has long been the understanding of these structures and their behaviors that are closely linked to their properties. One of the paramount properties of a crystal structure is its energy, as it is directly related to its stability and from which various other properties can be derived [1, 2, 3].

What defines the behavior of materials is indeed their crystal structures and the resulting properties. This behavior is defined in terms of responses to external influences. In real life applications, materials are selected based on their properties therefore, how they respond to the specific influences within the application [4]. Consequently, predicting crystal structure properties is vital in many fields that we may encounter in our daily lives. To achieve the best performance in a certain application, we must choose materials with characteristics that are specifically tailored for the application at hand. This requires the discovery of materials with the right properties, making property prediction an essential tool for innovation and efficiency across a wide range of sectors and applications.

In this thesis, we set out on a journey to the core of crystal structure energy prediction where the grace of crystallography meets cutting-edge machine learning techniques. In this context, data, algorithms, and this field's knowledge come together to unveil the profound mysteries held inside crystal structures.

Problem statement

In materials science, the two widely adopted conventional methods are recognized to be experimental measurement and computational simulation. Experimental measures, in addition

to their lengthy execution time, impose significant demands on tools and equipments, proper experimental environment, and the researcher's expertise, rendering the method inefficient [5]. Consequently, rather than undertaking laborious and costly experiments, materials scientists use quantum mechanics methods, as an alternative, to investigate and preliminarily assess novel materials. *Ab-initio* methods, such as Density Functional Theory in particular, are adopted to thoroughly understand materials properties through computer analysis, before engaging in physical synthesis and experimentation [6].

As opposed to experimental measures which take a period ranging from 10 to 20 years for a materials discovery (from study to first usage), computational methods can reduce this long, overwhelming time to 18 months [5]. Nevertheless, this reduced amount of time is still considered lengthy and not optimal. Therefore, a direct and alternate approach of accessing the relevant physical properties of crystal structure, without resorting to experimental measurements or computations, is unquestionably required.

In contrast with conventional methods, data-centered approaches use prior results to comprehend new situations. These data-centered approaches call for the use of Artificial Intelligence (AI) tools, more specifically, the machine learning (ML) process [6]. It typically uses algorithms designed to recognize and understand patterns within data, then use that learning to make predictions of new introduced data. The far-reaching applications of machine learning have been continuously evolving in different fields [7]. Indeed, the enormous amount of data being available has made the application of this set of statistical tools in many fields possible, or even essential [8], and the materials science field makes no exception [9, 5]. The problem to be addressed in this thesis involves the automatic learning of correlations existing between the collected data and the energy property, with the aim of making use of the acquired knowledge for the prediction of this property in potential new crystal structures.

Scope and limitations

In this study we focus on predicting the energy in crystal structures using machine learning-based techniques. While we aim for broad applicability, our work is limited to specific crystal systems and materials, facing the limitation and constraints of availability of crystallographic data and computational resources.

Our research tackles several key questions, including: how can machine learning approaches be employed to predict the energy property in crystal structures? What correla-

tions exist between the energy property and crystallographic features? Can we identify and validate a machine learning model that accurately predicts the energy property of crystal structures?

This thesis is expected to contribute to the field of crystal structure prediction by demonstrating the efficiency of machine learning-based techniques in predicting the energy property of crystalline materials. It will offer insights into the relationship and correlations between crystallographic features and energy, guiding the way to potential advancements in materials science and engineering.

Objectives and contributions

After a thorough analysis of the state of the art in crystal structure prediction, a clear idea of the challenges that still arise has been formulated. First and foremost, a critical challenge is to define the right crystal structure input for the learning process. This task is undeniably crucial since it has a relevant impact on the prediction outcome. Crystal structure databases (experimental or computational) present data under its raw form; in order to use this data for ML purposes, one has to first, extract numerical relevant information through a features engineering process. The result of such a process represents a suitable input for the modeling step. To this end, we have adopted two approaches for the representation of crystal structures: a structural approach and an atomic approach, both based on functions producing descriptors that are invariant with respect to rotation, reflection, permutation, and translation.

The second identified crystal structure prediction challenge is the modeling of the desired property. In this study, we investigate the modeling of the crystal structure energy property using machine learning-based techniques. Following the features engineering process, we propose two approaches to address this objective. The first one is a structural modeling approach which takes the structural descriptors as inputs and predicts the energy property by means of machine/deep learning algorithms. The second one is an atomic modeling of the relationship between the energy property and the atom-wise representation. We propose here a non-conventional deep neural network topology to support these descriptors.

In order to position this work within the context of the state of the art in crystal structure prediction, we assess the various implemented models to select the best-performing one and thereby validate it.

Thesis organization

The remainder of this thesis is organized into five chapters. The description of each chapter is presented hereafter.

The opening chapter of the present manuscript represents an intellectual pillar, offering a strong foundation for our investigation on crystal structure prediction within the materials science field. To ensure that readers from different backgrounds can seamlessly explore the next chapters, we carefully clarify the key concepts and terminology that underlie the area of research of this study. The main basic concepts carefully defined in this chapter are the fundamentals of crystal structures and their properties, a brief history of crystallography and quantum crystallography, as well as their role in crystal structure prediction.

The second chapter of this thesis provides a comprehensive overview of the ever evolving field of crystal structure prediction. It develops into two different yet linked sections, each examining an important aspect of the area. First, an analysis of crystal structure representation-related studies is conducted, revealing the distinct approaches adopted to model and represent crystalline matter. Then, we continue our survey by exploring crystal structure property prediction-related studies. This section covers recent cutting-edge machine learning-based approaches proposed to predict materials properties with impressive accuracy. These two sections together highlight the current state of the art and lay the ground for future innovative contributions.

The third chapter is where we initiate our own journey into the intricacies of crystal structure prediction. As a response to the first objective of our work, we delve deeper into the features engineering process of crystal structures in order to transform the raw collected data into powerful descriptors to fuel predictive models. Moreover, in the purpose of unveiling the nature of the relationship between dependent and independent data, the correlation between inputs and outputs of the resulting descriptors was investigated; thus, revealing more insights for the modeling step.

To address the previously mentioned second challenge, we present the fourth chapter intended for the modeling stage of our study. To this end, we proposed two approaches to model the relationship between the input data and the energy property. We investigated different ML algorithms for the first approach according to the input type and the correlation nature between the descriptors and the energy property to be predicted. In addition, we proposed a novel deep neural network topology that aligns with the data type of the second approach.

The last chapter is dedicated to the presentation of the results we obtained in crystal

structure prediction using the proposed machine learning-based solutions. As we navigate through a range of numerical data and graphical representations, the remarkable accuracy of our models is revealed. The evaluation of the models was performed using adequate strategies for proper analysis and rigorous assessment metrics. We further engage in deep discussions about the achieved energy property prediction results and their implications as related to the validation of the best performing model and the most robust descriptors.

We end this thesis with a general conclusion in which we provide a comprehensive overview of our main contributions, addressing each objective. Additionally, we set the stage for future advancements in crystal structure energy prediction as we present our perspectives.

Chapter 1

Crystal Structure Prediction Problematic Background

"Materials are probably more deep-seated in our culture than most of us realize."
– William D. Callister, Jr.

1.1 Introduction

Crystal structure prediction (CSP) has become a key area of research in the field of materials science. The quest for CSP was, and still is, mainly driven by the urgent need to create unique, novel materials with functional properties for applications spanning medicines, renewable energy technologies, and many other fields.

Our first chapter establishes the context before we get into the core of our research journey by examining the problematic background that motivates our study. The main objective of this chapter is to help the reader better grasp the contexts of the present study. For this purpose, the key elements of the crystal structure prediction field of research will be defined, starting from crystal structures and their fundamentals to their properties and a brief introduction to the history of crystallography as a classic, conventional CSP approach.

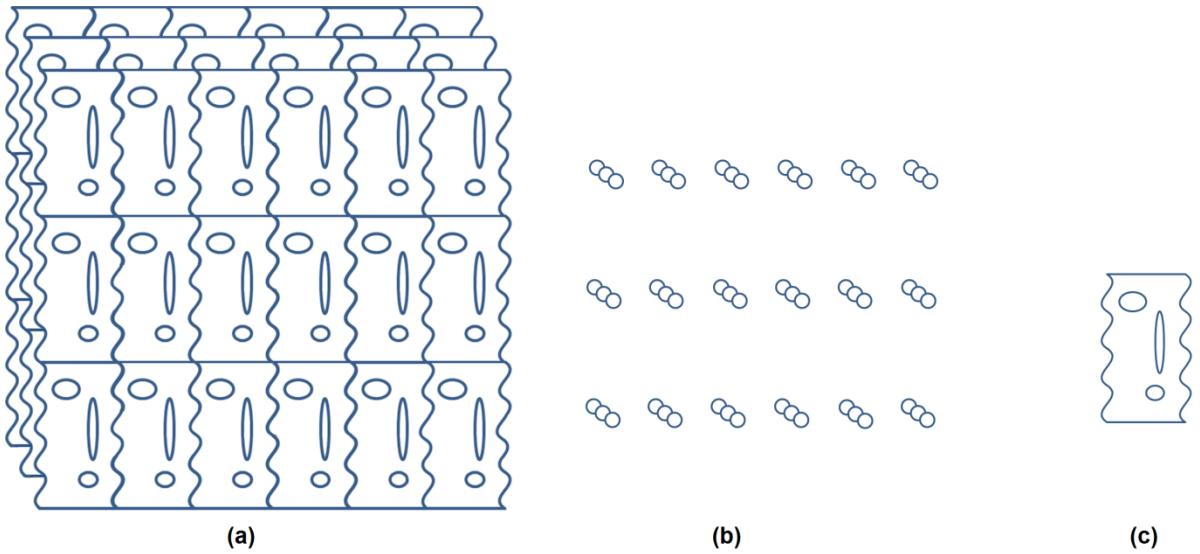


Figure 1.1: Representative depiction of the arrangement of atoms in a crystal structure. (a) Pattern, (b) arrangement, (c) motif [4].

1.2 Fundamentals of crystal structures

1.2.1 Crystal structures

A crystal structure is a solid material distinguished by the repetitive arrangement of its constituents. From a point of (Democritus's – Greek philosopher) view, a matter is composed of a unique combination of an infinite number of atoms. Long ago, a material was considered to be crystalline or amorphous according to whether it has a long range periodicity or a random arrangement of its atoms [10]. Later on, in the beginning of the nineteenth century, the matter's atomic nature and its details related to microscopic arrangements have become well established. Then, in 1869, Dmitri Ivanovich Mendeleev (Russian chemist) proposed the periodic table with all the elements arranged which is still acknowledged to be correct by now [10].

The particular regular arrangement of the atoms in a material is what defines a crystal structure. In order to understand this concept more intuitively, let's consider the drawing illustrated in Figure 1.1 (which itself does not represent a real crystal structure).

In Figure 1.1, we notice in (a) the repetitive appearance of the pattern (c) along the arrangement (b). Similarly, a crystal structure is composed of a motif (illustrated by c) that is translated from one point considered as the origin, to the other points (shown in b) to form a 3D structure as represented in (a).

1.2.2 Lattice and unit cell

If we project our previous example (described in Figure 1.1) on a material, a crystal structure can be defined by two elements: the arrangement that is referred to as "lattice" and the motif. Thus, in a crystal structure, the space lattice is a mathematical concept with a periodic pattern of points where each lattice point (node) is decorated with a motif [10, 4]. It is worth mentioning that the complexity of a material is not related to the lattice itself; a complex protein crystal structure and a simple pure metal one may share the same description with regards to lattice space, the motif however in each lattice point may range from 1 atom to thousands [11].

When we characterize the motif with geometrical parameters, we refer to it as "unit cell". It is defined as the simplest, smallest portion which is repeated in space. Since it is a three-dimensional space, we have three axes x, y, and z; the unit cell parameters are the length of the unit cell edges from the origin in the three directions noted as a, b, and c, and the angles between them labelled α, β, and γ such as [12]:

- a, b, and c are the lengths along x, y, and z axes, respectively
- α is the angle between b and c
- β is the angle between a and c
- γ is the angle between a and b

In order to visualize the above-mentioned parameters, we introduce Figure 1.2.

As we can observe, Figure 1.2 (a) represent a portion of a crystal lattice while Figure 1.2 (b) shows a body centered cubic (which will be discussed later in this chapter) unit cell with its parameters and atoms coordinates according to the illustrated origin.

If we take any lattice point (illustrated by orange circles in Figure 1.1 (a)) as the origin, we can jump to any other point in the crystal by translation. The translation vector (also called lattice vector) is defined as follows [4]:

$$t = ax + by + cz \tag{1.1}$$

There are four types of unit cells divided into two categories: primitive and centered, as explained in Figure 1.3.

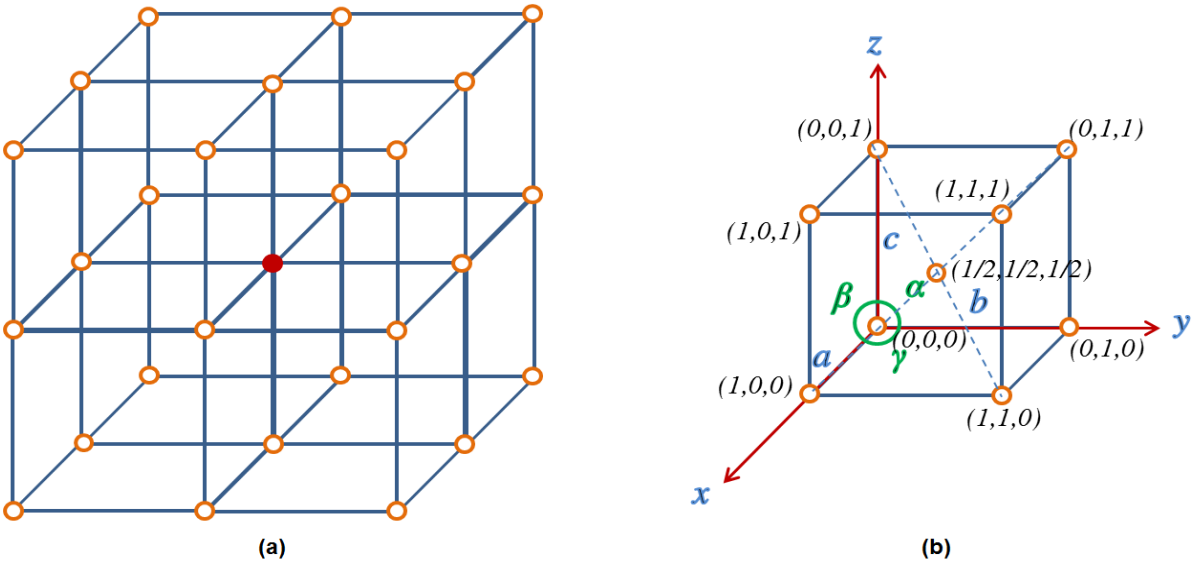


Figure 1.2: Schematic representation of a crystal lattice where (a) illustrates an example of a portion of the crystal lattice and (b) a cubic unit cell.

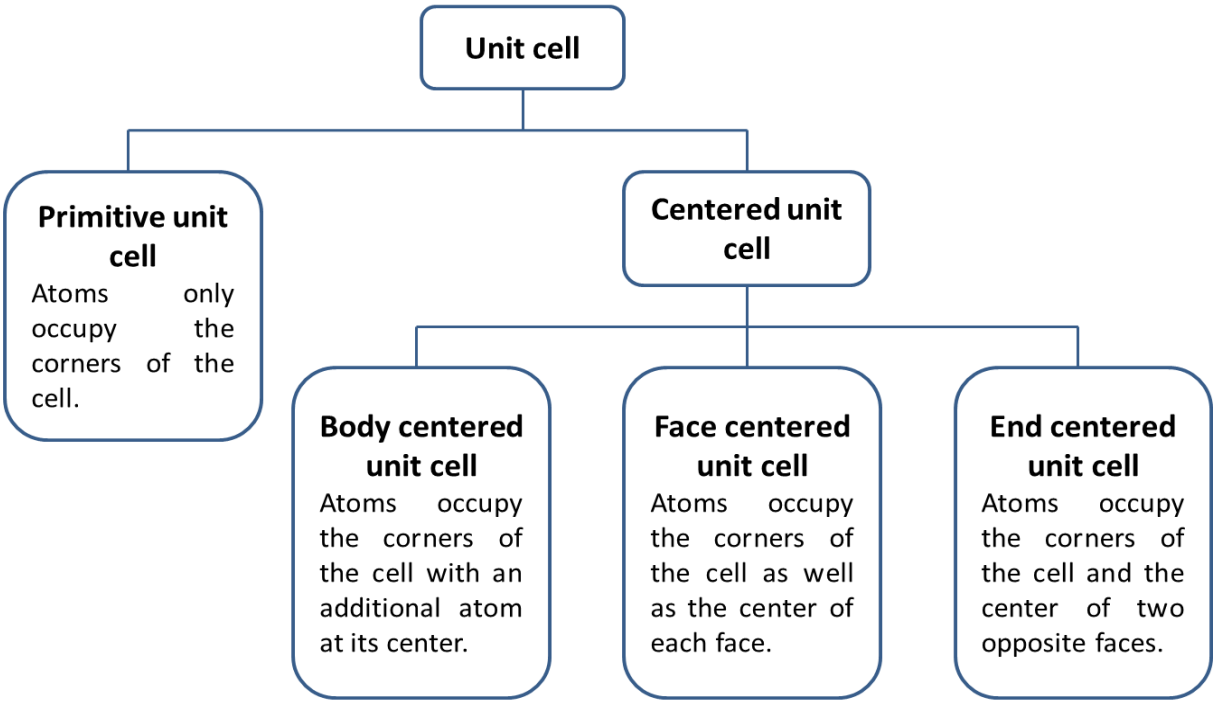


Figure 1.3: Types of unit cells and their categories.

1.2.3 Bravais lattices

Auguste Bravais has defined a number of lattice types to classify crystal structures. According to him, there are seven crystal systems designated as: Cubic, Tetragonal, Orthorhombic, Monoclinic, Triclinic, Trigonal, and Hexagonal. These seven crystal systems, along with the four types of unit cells make up twenty eight (28) lattice combinations; however, only fourteen (14) of those are possible called Bravais lattices from which we can build any crystal structure. Table 1.1 presents the different possible Bravais lattices and their parameters [10, 13].

All crystal systems share the feature of having a six-face shape except for the Hexagonal system having eight faces where the end faces have a six-side shape (hexagon), in this case the (x, y, z) axes are highlighted in red in Table 1.1.

The number of atoms (n) in a unit cell depends both on its type and its crystal system. In six-face shape crystal systems, the atoms at the corners of a primitive (p) unit cell make up one atom. Indeed, each corner atom is shared between eight adjacent unit cells (four on the top and four on the bottom) as illustrated in Figure 1.2 (a) by the red circle in the middle, thus:

$$n_p = 8 \times \frac{1}{8} = 1 \tag{1.2}$$

In an eight face shape crystal system (Hexagonal), the corner atoms make up two atoms since each corner atom is shared between six unit cells (three on the top and three on the bottom) in addition to the two atoms at the top and bottom center that are each shared between two unit cells (one above and one below). The Hexagonal primitive unit cell is not to be confused with an end centered one; even though it has two atoms each placed at the center of the end faces, these are in fact corner atoms shared between the 3 adjacent six-face shapes which constitute a single Hexagonal. Therefore:

$$n_p = 12 \times \frac{1}{6} + 2 \times \frac{1}{2} = 3 \tag{1.3}$$

If the unit cell is not primitive, the number of atoms becomes:

- Body centered (bc): $n_{bc} = n_p + 1$ (for the additional atom in the center)
- Face centered (fc): $n_{fc} = n_p + 6 \times \frac{1}{2} = n_p + 3$ (for the additional six atoms in the center of the six faces, each shared between two unit cells - case that does not appear in a hexagonal)

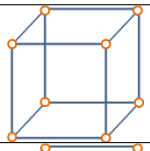
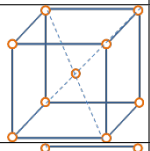
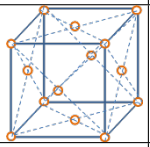
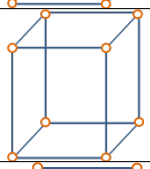
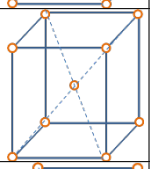
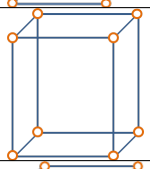
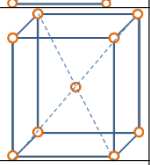
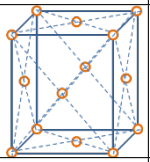
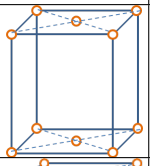
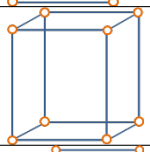
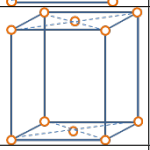
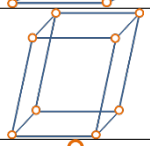
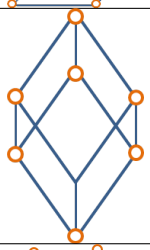
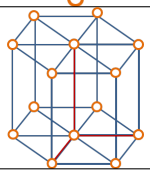
System	Primitive	Body centered	Face centered	End centered	Parameters
Cubic					$a = b = c$ $\alpha = \beta = \gamma = 90^0$
Tetragonal					$a = b \neq c$ $\alpha = \beta = \gamma = 90^0$
Orthorhombic					$a \neq b \neq c$ $\alpha = \beta = \gamma = 90^0$
Monoclinic					$a \neq b \neq c$ $\alpha = \beta = 90^0 \neq \gamma$
Triclinic					$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$
Trigonal					$a = b = c$ $\alpha = \beta = \gamma < 120^0$ $\neq 90^0$
Hexagonal					$a = b \neq c$ $\alpha = \beta = 90^0,$ $\gamma = 120^0$

Table 1.1: The fourteen (14) Bravais lattices with their unit cell parameters [14].

- End centered (ec): $n_{ec} = n_p + 2 \times \frac{1}{2} = n_p + 1$ (for the additional two atoms in the center of the end faces, each shared between two unit cells)

1.3 Crystallography

1.3.1 Birth of crystallography

Crystallography as a science has a long history which dates back to the 17th century. The crystals' exquisite symmetry has always raised the possibility of some sort of underlying order. Although only few humble experiments were conducted in this matter, it was obvious for scientists to claim that crystals must be composed of organized arrangements of tiny particles (known today as atoms and molecules) based on their symmetry and shape [15].

In 1895, the discovery of X-rays revolutionized the field of crystallography and was critically significant for its advancement. Two decades later, while scientists were debating whether or not X-rays were electromagnetic waves, a group of German physicists including Max von Laue, Paul Knipping, and Walter Friedrich were conducting experiments on X-rays and crystals. In 1912, two stunning discoveries were made by these physicists by beaming the X-rays through the crystals. Indeed, after the radiation scatter was captured on photographic plates, it was first confirmed that X-rays were in fact waves since they diffracted, thus settling a 17 years-old controversy, and secondly, this provided concrete proof of the atoms' underlying order in the shape of a lattice. This experimentation awarded Max von Laue a Nobel prize in 1914 [16].

In the summer of 1912, the physicist William Henry Bragg, after receiving an interesting letter describing Max von Laue's lecture, worked eagerly on X-ray diffraction in the University of Leeds, where he was a physics professor, with his son William Lawrence Bragg, a 22 years-old graduate student who happened to be on a holiday with his parents. Once returned to the University of Cambridge, W. L. Bragg had a sensational idea. He realized that, in addition to the existing component order of a crystal, its exact atoms arrangement could be deduced from the X-ray experiment initiated by Laue [17].

The Braggs demonstrated that the positions of atoms could be accurately determined by revealing the diamond's 3D crystal structure [15]. Moreover, the younger Bragg soon mathematically explained Laue's diffracted images which became known as Bragg's Law [17].

$$n\lambda = 2dsin\theta \tag{1.4}$$

The Equation 1.4 above illustrates how the X-rays wavelength λ, the interplanar separation d, and the angle of diffraction θ are related.

Not only their discovery has launched a brand-new scientific field named X-ray crystallography, father and son Braggs have also been awarded a Nobel prize in 1915 for their work, making William Lawrence Bragg, to this day, the youngest (scientific) Nobel prize recipient at the age of 25 [17].

1.3.2 Symmetry in crystal structures

A crystal structure possesses a symmetry that allows one to interchange a part of it with another while this material remains unmodified. A symmetry is defined by two items: the symmetry operation and the symmetry element. The former is an action performed on the body with respect to the latter such that the before and after positions of the body are indistinguishable. The geometrical object representing the symmetry element may vary between a point, a plane, or an axis [12].

If we omit the identity operator, there are four main symmetry operations defined as follows:

Translation. It includes moving the crystal in a way that each atom is replaced by a neighbor that is identical. As previously discussed in the definition of a crystal structure and unit cells, a translation symmetry operation is characterized by a translation vector considered as the symmetry element. It is noteworthy that only in an infinite solid can a translation be considered a real symmetry operation [4].

Rotation. This symmetry operation causes the crystal to revolve around a symmetry axis, representing the symmetry element, that runs through the crystal. It is characterized by the angle and the direction of the rotation which is positive for counterclockwise. The rotation angle is most commonly written as a fraction ($\frac{2\pi}{n}$) with n representing the order of the rotation. Theoretically, the value of n ranges between one (1) and infinity (∞) [4]. Figure 1.4 presents examples of rotation symmetry.

Figure 1.4 (a) and (b) illustrate two rotation symmetry operations of order three and six, respectively; while Figure 1.5 (a) describes a realistic example of a material with a rotation symmetry where we can clearly see how the rotation axis (element) passes through it.

Reflection. This operation is the most intuitive since we daily encounter a real example of it while using a mirror. In a solid, a reflection symmetry operation swaps out the crystal's parts on each side of a symmetry plane element referred to as mirror plane [4]. Figure 1.5 (b) reproduces the same realistic rotation example but with a reflection operation.

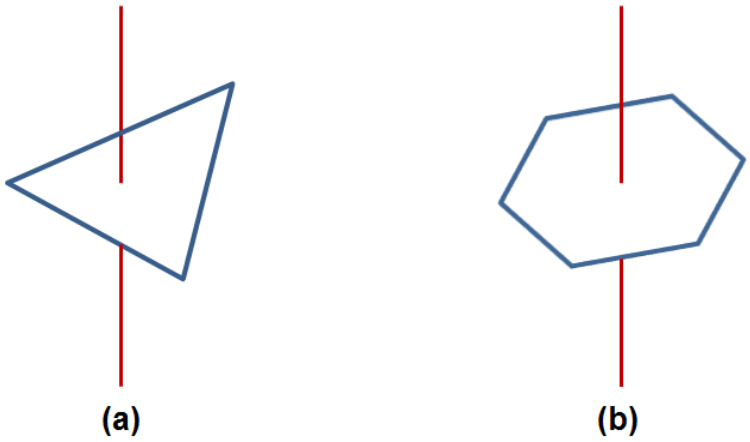


Figure 1.4: Examples of a rotation symmetry where (a) $n = 3$ and (b) $n = 6$ [4].

Inversion. Every atom is moved by inversion to a different location such as the before and after positions of the atom are lined up. The center of the lines holding as ends the old and new locations of atoms is the center of the inversion which is the element of the symmetry. Figure 1.5 (c) is an example of an inversion operation with its element highlighted in red.

In addition to four above-mentioned symmetry operations, we identify two more known as improper rotation: roto-reflection and roto-inversion. The former represents a rotation operation followed by a reflection one, and the latter a rotation operation followed by an inversion.

In addition to the fact that a crystal structure belongs to one of the fourteen Bravais lattices, and therefore, one of the four unit cell types and one of the seven crystal systems, it also can be categorized according to its symmetry. Bravais lattices can be classified into 32 different crystal classes referred to as point groups each corresponding to a certain possible combination of inversions, reflections, and rotations (pure and improper). If we include the translation symmetry, it is mathematically possible to produce 230 different arrangements of atoms in a periodic pattern; these are called space groups [10].

1.3.3 Quantum crystallography

Following the success of X-ray crystallography, theorists began to activate in this field using quantum mechanics. However, this task turned out to be far too complicated. Indeed, in quantum mechanics, in order to compute the crystal structure or the properties of a given system, we need to get all the information which is provided by the electron wavefunction. For this matter, we need to solve the Schrodinger equation of that system. The main problem of

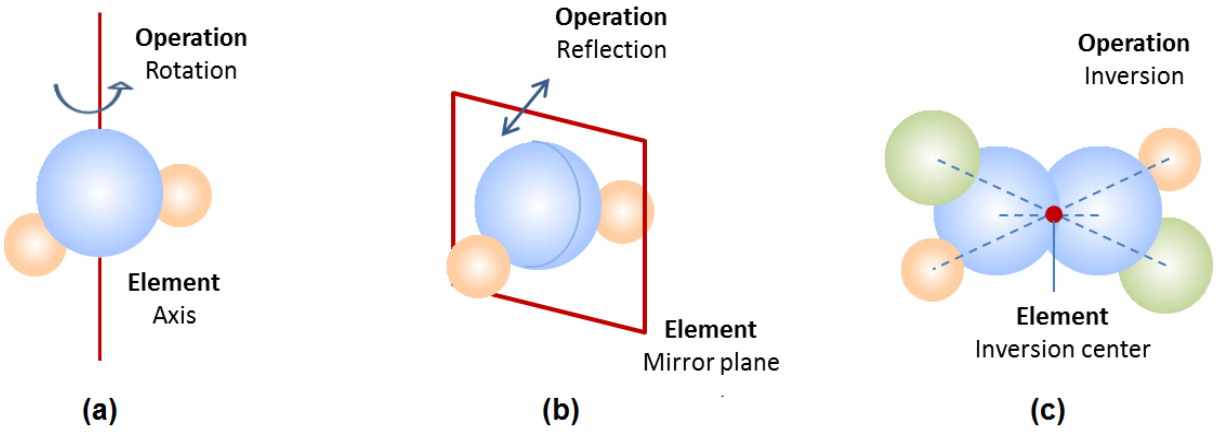


Figure 1.5: Examples of three symmetry operations with their elements, including (a) a rotation and its axis, (b) a reflection and its plane, and (c) an inversion and its center [4].

this equation is its complexity; it increases enormously with the system's constituent particle number. A crystal typically has 10^{25} valence electrons, which interact with each other; it requires a simultaneous resolution of more than 10^{25} variables. Thus, an approximation of the Schrodinger equation's solution of a many-body system is strongly needed [18].

It all started with the Schrodinger equation below:

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi \tag{1.5}$$

With i, $\hbar$, ψ, and $\hat{H}$ representing the imaginary number, a constant, the wave function and the Hamiltonian, respectively. Schrodinger states that the total energy is the sum of the potential energy V and the kinetic energy K (Equation 1.6) [19].

$$E = V + K \tag{1.6}$$

Where:

$$V(\vec{r}) = -\frac{e^2}{\vec{r}}, K = \frac{\left|\vec{P}\right|^2}{2m}, \vec{P} = \frac{\hbar}{i}\nabla, \nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}) \tag{1.7}$$

With e, r, P, m, and ∇ being the charge, the distance, the momentum, the mass, and the gradient, respectively.

In 1927, Thomas and Fermi described a model for calculating atomic properties which was purely based on the electron density. Their model is an approximate method for finding the electronic structure of atoms using just the one electron ground-state density; as for the

kinetic energy, they adopted a local density (LD) approximation. However, this model had some severe deficiencies because of its poor description of the outer regions of an atom and was too crude to bind molecules [20].

In the same year (1927), Born-Oppenheimer approximation was introduced; it assumes that the motion of atomic nuclei and electrons in a molecule can be separated since the nuclei havs much slower motion than electrons due to their mass difference, meaning that it allows the wavefunction of a molecule to be broken into its electronic and nuclear components [21].

$$\hat{H}\psi = E\psi \tag{1.8}$$

$$\hat{H} = -\sum_{i=1}^{N_e} \frac{\hbar^2}{2m}\nabla_i^2 - \sum_{i}^{N_n}\sum_{j}^{N_e} \frac{e^2 Z_i}{\left|\vec{r_j} - \vec{R_i}\right|} + \sum_{i<j}^{N_n} \frac{e^2 Z_i Z_j}{\left|\vec{R_i} - \vec{R_j}\right|} + \sum_{i<j}^{N_e} \frac{e^2}{|\vec{r_i} - \vec{r_j}|} \tag{1.9}$$

With N_e electrons at positions r_i and N_n nuclei at positions R_i and charge Z_i.

As we can see in the Born-Oppenheimer Hamiltonian (Equation 1.9), there are four terms, the kinetic energy (nuclei considered as immobile), the electron-nuclei interaction, the interaction between nuclei and the interaction between electrons [21].

A year later (1928), Dirac added an exchange energy functional term to the Thomas-Fermi model. The model (TFD) is however inaccurate because the representation of the kinetic energy functional term is just an approximation and the electron correlation effect is completely neglected [22].

The same year in 1928, the Hartree-Fock (HF) method was introduced as an approximate solution to the Schrodinger equation. The equations in this method are obtained by varying one-electron wavefunctions [21]. The HF, despite being far more beneficial than TFD, is still insufficiently precise for the prediction of energy in chemistry because of the underestimation of bond energies [23].

In 1951, Slater proposed the so-called Hartree-Fock-Slater approximation as a simplification of the Hartree-Fock method by combining it with Thomas and Fermi's theory [20].

1964 is considered to be the birth year of density functional theory (DFT) where Hohenberg and Kohn proposed to substitute the complicated many-electron wavefunction containing $3N_e$ variables, with the functional of electron density, which only contains 3 spacial variables. Two theorems were introduced stating that 1) the ground state properties of a many electron system depend only on the electronic density [21, 23] and 2) the correct ground state (GS) density is the one that minimizes the total energy (E) (Equation 1.10) [24, 22].

$$E_{GS} = minE[n, V_{ext}] \tag{1.10}$$

Where n is the electronic density, and V_{ext} is the external potential.

Kohn and Sham introduced a year later the concept of a system of non-interacting particles moving in an external potential [25]. Ever since and until today, DFT has become an inevitable tool in most branches of chemistry and solid state physics.

1.3.4 Density functional theory

Density functional theory is in principle an exact theory to describe the electronic structure because it's purely based on the electron density distribution, instead of the many-electron wave function. Indeed, Density Functional formalism shows that ground state and other properties of a system of electrons in an external field can be determined just by knowledge of the electron density distribution [22].

In quantum mechanics, the electron density is defined by the probability measure that an electron occupies a very small space surrounding a certain point. It is a scalar quantity depending upon three spatial variables [26]. Whereas, a functional is simply a function that depends on another function. Since the starting point of DFT, many researchers focused on functionals. Today, a lot of different functionals exist; these functionals can all be grouped into four main families, namely: LDA, GGA, meta-GGA, and hybrid functionals [27].

1.3.4.1 Local Density Approximation (LDA)

In physics, the most widely used approximation is the local-density approximation, where the functional depends only on the local density at a given point, that is the coordinate where the functional is evaluated. This means that the exchange-correlation energy density at every position in space for the molecule is the same. In LDA, the exchange–correlation energy is typically separated into the exchange part and the correlation part (Equation 1.11) [28].

$$E_{XC}^{LDA}[n] = E_{X}^{LDA}[n] + E_{C}^{LDA}[n] \tag{1.11}$$

Since LDA assumes density is the same everywhere, it has a tendency to underestimate the exchange energy and over-estimate the correlation energy [28]. The errors due to the exchange and correlation parts tend to compensate each other to a certain degree. But in

order to correct this tendency, it is more common to expand in terms of the gradient of the density [29].

The performance of LDA on structural, elastic, and vibrational properties is considered to be good enough. However, there's an overbinding problem with binding energies. In addition, the activation energies in chemical reactions are unreliable and the relative stability of crystal bulk phases can be uncertain [29, 28].

1.3.4.2 Generalized Gradient Approximation (GGA)

The generalized Gradient Approximation's functional is a way to improve the accuracy provided by the LDA one. It depends, not only on the local density, but also on its gradient. In fact, Most of GGA functionals are constructed in the form of a correction term which is added to the LDA functional (illustrated by the second term of Equation 1.12) [28].

$$E_{XC}^{GGA}[n] = E_{XC}^{LDA}[n] + \Delta E_{EX}\left[\frac{|\nabla n(r)|}{n^{\frac{4}{3}}(r)}\right] \tag{1.12}$$

GGA functionals successfully corrected the overbinding problem of the LDA ones and improved both the activation energies in chemical reactions and relative stability of crystal bulk phases' description. However, GGA's workfunctions for several metals turn out to be somewhat smaller than in LDA, and more importantly, Van der Waals forces are not included (which is a major limitation especially for applications in chemical field and it is not universally acceptable) [30].

1.3.4.3 Meta-GGA

Potentially more accurate than the GGA functionals are the meta-GGA functionals. This accuracy comes from the fact that they include the second derivative of the electron density which is the Laplacian $\nabla^2 n(r)$ whereas GGA includes only the density and its first derivative in the exchange–correlation potential [31].

In practice, instead of using the Laplacian in Meta-GGA, one usually includes the kinetic energy density (Equation 1.13) since it is more stable numerically [32].

$$\tau(r) = \frac{1}{2}\sum_i |\nabla \phi_i(r)|^2 \tag{1.13}$$

1.3.4.4 Hybrid functionals

Hybrid functionals Include fractions of exact Hartree-Fock exchange energy, calculated as a functional of the Kohn-Sham molecular orbitals as illustrated in Equation (1.14) where the first term represents an LDA or GGA exchange correlation, and the second is the Hartree-Fock exchange [33].

$$E_{XC} = (1 - a)E_{XC}^{DFT} + a_X^{HF} \tag{1.14}$$

As a brief comparison between the different functionals, LDA functionals are the simplest of the 4 families; however, they have the least accuracy. On the other hand, hybrid functionals are the most accurate but the least simple. The GGA and meta-GGA functionals are somehow in between having and average simplicity-accuracy ratio.

1.4 Formation of crystal structures

As previously explained, a crystal structure is defined by the arrangement of its atoms periodically in a crystal lattice. Atoms are stacked tightly together in result of their chemical reactions due to the attractive force between atomic nuclei and electrons. This atomic bonding constitutes the essence of the formation of crystal structures. The type of the atomic bonding determines the sort of interaction an atom has with its neighbors according to which an atom might exist in several energy states [11]. Moreover, we should emphasize that the potential energy is a key element in crystals since their formation is directly related to the utterly ordered state of minimal potential energy [34].

The atomic bonding responsible for the formation of crystal structures can be categorized into two main classes. The first one (and most commonly encountered) comprises ionic, covalent, and metallic bonding where ionic and covalent bonding are the strongest. The second less common class which has a weak force of attraction includes hydrogen and van der Waals type of bonds [11].

Ionic bonding. In this type of bonding, two or more atoms attract each other in a way that one loses one or more electrons for others to gain them. The formed molecule is neutrally charged and each atom has the configuration of a noble gas. For example, in the NaCl molecule, the Sodium atom (Na) loses one electron in order to reach the Neon (Ne) noble gas configuration for the Chlorine (Cl) atom to gain it so that it reaches Argon (Ar) configuration. The balance of the attracting and repellent electrostatic forces forms

the foundation of the bond. In a crystal, ions are arranged in a non-directional way that produces a macroscopically neutral material [11].

Covalent bonding. Atoms in this case bond by sharing electrons in a manner that the molecule they form is neutrally charged and each atom has one of the noble gases' configuration. For instance, the Oxygen atoms in the O_2 molecule each shares two electrons; with that bonding, they each reach the configuration of the Neon (Ne) noble gas and the total charge of the O_2 molecule is neutral. In crystals or molecules, the electrons that are shared are within the direct line between the atoms which is caused by the density of electrons that are concentrated between the nuclei. The interactions caused by a covalent bonding can be, in organic compounds, of saturated or unsaturated nature: saturated bond is basically a single bond while unsaturated bond contain one or more double bonds or even a triple bonds [11].

Metallic bonding. Unlike ionic and covalent bonds which are chemical-valence-based, a metallic bond is considered as force that holds metal ions together forming an electrostatic force. The force of attraction acts between positive ions and electrons of either identical or different atoms as observed in many alloy structures formation. The metallic bonds typically act between atoms and their eight or twelve first neighbors, reaching, thus, a more stable configuration by sharing their outer shell electrons [11].

Hydrogen bonding. Also referred to as H-bond, is a type of intermolecular force occurring between hydrogen-based molecules. It results from the attractive force between a hydrogen atom covalently bonded to a very electronegative atom such as a nitrogen, oxygen or fluorine atom and very electronegative atom. H_2O (water) molecules for example are bonded through hydrogen bonding [35].

Van der Waals forces. An attractive force between close neutral molecules due to polarization resulting from instantaneous redistribution of charges. Even when other bonds (ionic, covalent, and metallic) are absent, this electrostatic attraction / repulsion, called Van der Waals force, can be present between any two molecules [4]. In other words, when a molecule has a dipole (separated charges resulting from an uneven electron distribution), it possesses two ends one of which is positive and the other negative. Like magnets, these poles attract and repel opposite charges and like ones, respectively [13].

1.5 Materials properties

The essence of the materials science field is to study materials properties. Such studies must indeed be conducted at the atomic scale. In fact, what determines the properties of a material are the atoms it is composed of and the types of bonds that connect them. First, among the features of an atom, the shape of it is an important aspect in this context [10]. In addition to that, the distribution and arrangement of atoms make up the uniqueness of a material [4]. These elements are undeniably crucial to understand the properties of a structure. Second, there is a direct relationship between the atomic bonds of a material and its behavior. If we take the covalent bond as an example, as it hooks atoms very firmly, it makes up strong materials when there is not another weaker intermolecular force binding molecules. For instance, diamond, which is known to be the hardest material, is composed of carbon atoms all bonded together through covalent bonding. Water however is not a strong material since, even though Hydrogen is covalently bonded to oxygen (intramolecular), the H_2O molecules are bonded through hydrogen intermolecular forces which are much weaker and keep braking and reforming. While materials with covalent bonds are characterized by insulative or semi-conductive property [11], ionic crystals, whether dissolved in water or in a molten state, conduct electricity because of the dissociation of the crystal ions. Ions, once free, move to positive poles and negative ones carrying the electrical charge. Moreover, as strong bonds typically make strong/hard solids, the weaker attraction forces make materials with corresponding characteristics. As in case of graphite, the softness and lubricating properties are caused by the weak Van der Waals forces [4]. These intermolecular forces, although weak, are not any less important; in fact, without them, life would not exist as we know it. After all, it is the hydrogen bond that is responsible for the life-sustaining qualities of water and protein and DNA structure stabilization.

Crystal structure properties define the behavior of the material under certain conditions. In other words, when a material is exposed to an external influence, it has a particular response. The external field might be temperature, electricity, pressure, gravity, magnetism, etc. An example of a material's response is the deflection of a steel beam when an external load is applied to its ends, or the induction of an electrical field in a conductor when it is moved through an external magnetic field.

In real life applications, materials are chosen according to their properties and thus, how they behave as a response to the influence present in the application [4]. In industry for instance, plastic is used for different appropriate characteristics; we find it in toy production for example since it has the capacity to get molded into various shapes, it is light in weight,

and it is strong enough to handle small pressures. Similarly, tar is used to coat the ground and is supposed to hold regardless of the temperature, weather, and vehicles driven over it.

Lately, a great deal of materials science studies is dedicated to determine crystal structure properties for specific applications, and thus, identify which materials perform better given a certain application. This has been the case of many industries in order to have more efficiency and less potential cost. As in the pharmaceutical industry, drug discovery has taken a major role in determining materials with desirable therapeutic properties. In the renewable industry, we find that semiconductor materials are adequate for the conversion of solar energy to electricity; in addition to that, nanomaterials with their large surface properties are very useful in terms of efficient light absorption [36]. Likewise, materials discovery in the electronics industry makes no exception in studying solids with desired properties for particular applications in many fields like; dielectrics [37], ferroelectrics [38], oxides ion-conducting [39], piezoelectrics [40], pyroelectrics [41], photocatalytics [42], multiferroionics [43], microelectromechanical (MEM) devices [42], humidity sensors [43], water purification (due to the photocatalytic property), spintronics for many state memories [43], etc.

1.6 Conclusion

In closing, this chapter has laid the stage for our investigation into crystal structure prediction. It has emphasized the importance of our research with a focus on crystallography's rich history. The essential aspects of crystal structure prediction were explained for a better understanding of this study's field of research. This simple, comprehensive overview of crystal structure prediction ensures that the subsequent chapters are accessible and unambiguous.

In the next chapter, we will explore the pressing contemporary challenges of crystal structure prediction and how the most relevant works of the state of the art addressed them.

Chapter 2

State of the Art on Crystal Structure Prediction

> *"One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition."* - John Maddox.

2.1 Introduction

In the field of materials science, the pursuit to predict and understand crystal structures has quickly evolved. This is due to the extensive importance that researchers and scientists accorded to this discipline, leading them to dive deeper into the treacherous waters of crystal structure prediction. Here, we set out on a quest to investigate the most recent developments, approaches, and breakthroughs in the area of crystal structure prediction as we seek to situate our work within this dynamic and ever-evolving scientific field.

As a start, we define and specify the problem and challenges that our study seeks to solve. Then, we proceed to the literature review to explore both the representation of crystal structures and their prediction.

2.2 Problem statement

The modeling of the relationship between a structure and its properties illustrated by a behavior or an activity is extremely important in the scientific field of materials science. In addition, it could greatly impact the progress and the improvement of many technological

fields [5]. There has been a lot of interest in crystal structure property prediction across disciplines, as it has applications in many different fields. Indeed, crystal structure prediction's major role is to assist researchers in finding particular stable compounds characterized by desirable and application-suitable properties before synthesis in the lab. In industry, competing companies can make use of crystal structure prediction to either protect their own patents or even break patents of other companies [44].

In materials science, experimental measurement and computational simulation are known to be the two widely used conventional methods [5]. One way to describe crystal structures is powder X-ray diffraction (referred to as XRD). This representation as raw data source is however intricate since the 3D distribution of electron-density becomes 1D powder diffraction pattern. The crystal symmetry of many low-symmetry phases cannot be accurately determined from a powder XRD pattern because of this complication [45].

On the other side, experimental measurement is a simple and intuitive approach of materials research. It typically involves the analysis of microstructure and property, synthetic experiments, property measurement ...etc. This method is nonetheless carried out in an ineffective manner over a lengthy period of time and it places great demands on the tools, the setting for the experiment, and the researcher's skill and expertise [5]. Indeed, the time intensity process of discovering and characterizing new materials is justified by the fact that untested compounds must be synthesized under a lot of trial-and-error settings, and certain chemical reactions might take days to weeks to complete. Plus, numerous untested materials include pricey exotic compounds or elements. Then, samples must be characterized and analyzed for crystal structure and microstructure, which adds to the reagents cost [46].

Considering the seven stages of a new material search from the discovery to the manufacturing and deployment, passing by development, optimization of properties, design and integration of the system, and the certification, the amount of time it takes to find new materials is astoundingly long, usually 10 to 20 years from start study to first usage. Not to mention that each one of these stages requires distinguished researchers expertise and/or expensive elements and reagents [5].

Alternatively, materials scientists, physicists, and chemists increasingly use ab initio (first principle of quantum mechanics) approaches to anticipate the characteristics of materials using the basic quantum mechanical equations. Instead of conducting time-consuming/expensive experiments, these technologies enable scientists to examine and predict novel materials, and in certain situations, to even suggest brand-new and improved materials [6]. The idea behind the usage of ab initio methods, such as DFT, is to fully understand the characteristics of ma-

terials by computing before physical synthesis and testing. Consequently, they have become commonplace instruments in the field of materials science [47]. The advantage of the computational approaches compared to the experimental process is that there is no need for the costly experimental aforementioned environment, and the needed period can be shortened from 10 to 20 years, as determined by conventional procedures, to 18 months [5].

Crystal structure prediction can be separated into two main sub-problems, namely the search problem and the ranking problem [48]. The former is represented by a sampling method for the configuration space. This procedure is quite challenging considering the fact that the number of possibilities to arrange atoms in space is gigantic. Given a unit cell with N atoms, the corresponding number of possible structures is $c \sim exp(ad)$, where a and d represent system-specific constant and the degrees of freedom number, respectively, with $d = 3N + 3$ in case all none of the N atoms' locations are correlated. To significantly simplify this problem, a relaxation process is introduced. Although the complexity of the problem would still remain exponential, the number of possible configurations is largely decreased when each generated structure goes through relaxation by bringing it to a local energy minimum. The second sub-problem consists in accurately measuring the structural energies. However, appropriately ordering structures by energy is very arduous since the energy differences between various polymorphs are frequently quite minor [49].

Unfortunately, due to the inherent limits of both experimental and theoretical approaches, it is challenging to employ either of these two methodologies to speed up the materials discovery and design process [5]. In addition to the previously mentioned drawbacks of experimental methods, Quantum mechanical methods such as DFT come with a significant computational cost. The complexity of its calculations increases cubically with the atoms number, and these calculations are repeatedly performed throughout structural relaxation [48].

Consequently, there is an undeniable need for a direct and alternative method to access the physical property of interest without having to solve the Kohn-Sham density functional theory (KS-DFT) equations [50]. As opposed to quantum mechanical methods, which do not incorporate earlier computations when examining a new system, data-centered methods make use of previous results to understand novel situations [6].

One way to do that is by codifying information gathered from prior experience (computational or experimental) to make automatic informed estimates about compounds that are likely to arise in a new unknown system [47]. An appealing option of this kind is provided by machine learning approaches. After the ML model has been trained on a representative

training set of crystal structures, ML-based computations are extremely quick, generally able to predict a particular material's properties in fractions of a second [50].

2.3 Literature review on crystal structure prediction with machine learning

Machine learning is a powerful tool considered as an effective substitute to the time consuming classic quantum mechanical operations. With the huge number of data that is nowadays available, it has become almost necessary to lean towards automatic approaches. Although many works have focused on machine learning to solve materials science prediction problems, this field still presents many challenges. The two main tracks of crystal structure prediction via machine learning are 1) ML-suitable crystal structure data representation, and 2) ML-based modeling and optimization of crystal structure prediction [51].

2.3.1 Representation of crystal structures

This task is undeniably crucial since it has a relevant impact on the prediction outcome. Computational and experimental databases present data under its raw form; in order to use this data for ML purposes, one has to first extract numerical relevant information through a features engineering process. The result of such a process represents a suitable input for the modeling step [51, 52].

The Bravais matrix with the system's coordinate is not enough to represent a system; in fact, it is also not suitable for a learning process. Indeed, for a single crystal structure, there is an infinite number of representations that the computer would treat as distinct materials because of the symmetry operations of crystal structures [50].

The representation of crystal structure features is most commonly referred to as "descriptors" [53]. Below are the most recognized descriptors in the field of materials science.

2.3.1.1 Coulomb matrix

It is a straightforward global descriptor that replicates nuclear electrostatic interaction. It is represented through an $n \times n$ matrix, where n is the number of atoms. It was first introduced by M. Rupp et al. [53] to represent and describe molecules. The Coulomb matrix is defined as follows:

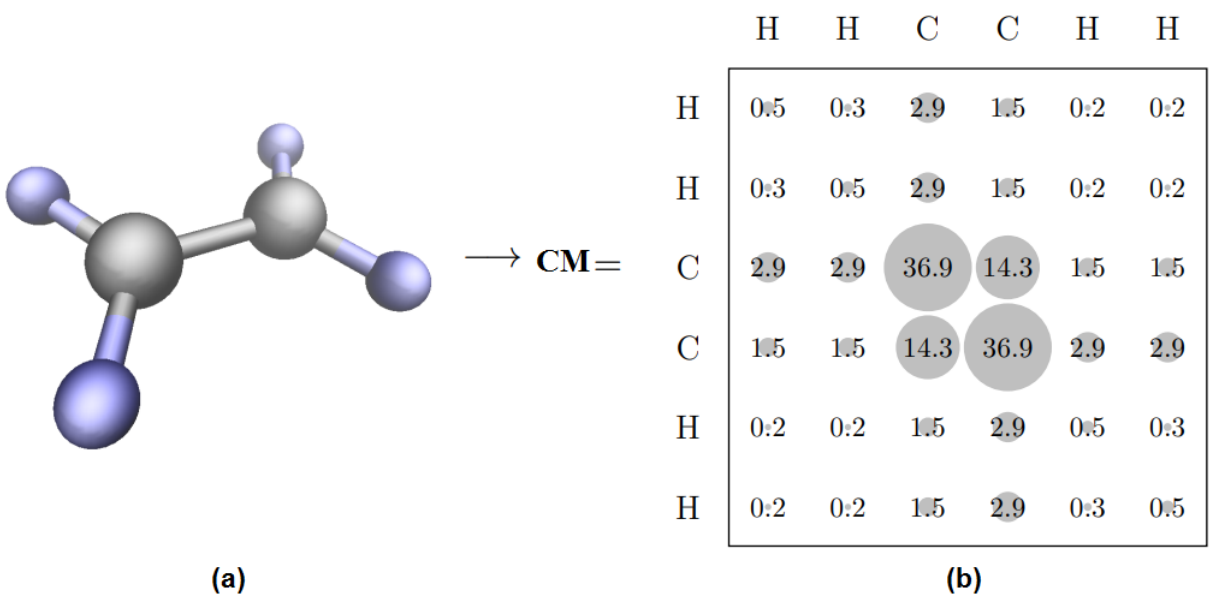


Figure 2.1: Coulomb matrix representation of C_2H_4 molecule [54].

$$x_{ij} = \begin{cases} 0.5Z_i^{2.4} & for i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & for i \neq j \end{cases} \tag{2.1}$$

Where Z and R represent the nuclear charges and Cartesian coordinates, respectively.

Figure 2.1 is an illustration of the C_2H_4 molecule represented through a Coulomb matrix.

Unfortunately, there is no atom ordering definition in the Coulomb matrix; which means that by permuting atoms in matrix's rows and columns, one would end up with many different Coulomb matrices for the same molecule. Authors in [54] have fixed this issue by introducing "dummy atoms" to get a well-defined atom ordering resulting in a unique Coulomb matrix for each molecule. However, this representation is not suitable for learning purpose since, given a database with materials having different number of atoms, it would result in Coulomb matrices with different dimensionalities [54]. Moreover, Coulomb matrix representation cannot be used to directly describe infinite periodic crystals [50].

F. Faber et al. proposed in [55] a generalization of the Coulomb matrix to overcome its infinite periodic description problem. Indeed, to adapt the Coulomb matrix representation on crystal structures, the three following generalizations were introduced:

A. Ewald sum-based generalization. A matrix in which every element is connected to the Ewald sum of two distinct atoms' electrostatic interactions that is repeated throughout the lattice in the unit cell. An element of Ewald sum-based Coulomb matrix is defined as

follows:

$$x_{ij} = x_{ij}^{(r)} + x_{ij}^{(m)} + x_{ij}^{0} \tag{2.2}$$

Where the three terms represent the real space calculations of short range interaction, the reciprocal space-calculated interaction of long range, and a constant, as defined in Equations 2.3, 2.4, and 2.5, respectively.

$$x_{ij}^{(r)} = Z_i Zj \sum_{L} \frac{erfc(a \left\| r_i - r_j + L \right\|_2)}{\left\| r_i - r_j + L \right\|_2}, (i \neq j) \tag{2.3}$$

$$x_{ij}^{(m)} = \frac{Z_i Z_j}{\pi V} \sum_{G} \frac{\frac{e_2^{-\|G\|_2}}{(2a)^2}}{\left\| G \right\|_2} cos(G \cdot (r_i - r_j)), (i \neq j) \tag{2.4}$$

$$x_{ij}^{0} = -(Z_i^2 + Z_j^2)\frac{a}{\sqrt{\pi}} - (Z_i + Z_j)^2 \frac{\pi}{2Va^2}, (i \neq j) \tag{2.5}$$

Where L and a illustrate the lattice vectors and length parameter, respectively, in Equation 2.3, and G and V represent the lattice vectors and the volume of the unit cell, respectively, in Equation 2.4.

B. Extended Coulomb matrix generalization. An extension of the representation matrix size of the traditional Coulomb matrix by considering the neighboring unit cell number. This form has the advantage that it is easier to evaluate than the Ewald sum matrix.

C. Sine matrix. A simplified matrix that uses a sine function of the atoms' crystal coordinates to imitate the periodicity and fundamental characteristics of the elements in the Ewald sum matrix. The matrix's elements are defined as:

$$x_{ij} = \begin{cases} 0.5 Z_i^{2.4} & for i = j \\ Z_i Z_j \tilde{\Phi}(r_i, r_j) & for i \neq j \end{cases} \tag{2.6}$$

$$\tilde{\Phi}(r_i, r_j) = \left\| B \cdot \sum_{k=x,y,z} \hat{e}_k sin^2[\pi \hat{e}_k B^{-1} \cdot (r_1 - r_2)] \right\|_2^{-1} \tag{2.7}$$

Where B and $\hat{e}_x, \hat{e}_y, \hat{e}_z$ represent the lattice vectors-based matrix and the unit vectors coordinates, respectively.

To further prove the effectiveness of these descriptors, authors in [55] conducted a formation energy prediction using each of the three proposed representation as inputs. For the modeling stage, authors opted for kernel ridge regression (KRR) to train and test data from the Materials Project (MP) database [56]. The proposed descriptors turned out to be

suitable for both crystal structure representation and learning purpose where the sine matrix was acknowledged as the most effective.

2.3.1.2 The partial radial distribution function (PRDF)

PRDF [50] is a material representation which, for each pair of atom type, the pair-wise distance distribution is examined. If we consider an atom of type α as the center, and an atom of type β in a shell of radius and width of r and dr, respectively, the PRDF representation is given by Equation 2.8, as averaged over an atom type.

$$g_{\alpha\beta}(r) = \frac{1}{N_\alpha V_r} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \theta(d_{\alpha_i\beta_j} - r)\theta(r + dr - d_{\alpha_i\beta_j}) \tag{2.8}$$

Authors in [50] introduced this materials representation and validated it through a case study of density of electronic states (DOS) prediction at the Fermi energy. For this purpose, a training set was generated using DFT with an LSDA (Local Spin Density Approximation) functional. The ML models used were KRR implemented as linear-based, Gaussian-based, and Laplacian-based which was trained on the generated data and tested on training-independent data composed of ICSD [57] (Inorganic Crystal Structure Database)-selected samples, metallic alloys (PbAl), and a CBN (Carbon, Boron, and Nitrogen) solid solution, in order to validate the study.

2.3.1.3 Elemental and structural descriptors

The work cited in [58] proposed a set of features to represent and describe crystal structures as well as molecular systems for machine learning. This descriptor represents a compound ξ through a matrix (Equation 2.9) of size $N_a^{(\xi)} \times N_x$, i.e. the number of atoms by the number of features describing each atom, with $N_x = N_{x(ele)} + N_{x(st)}$ where the first and second terms represent elemental and structural features, respectively.

$$x^{(\xi)} = \begin{pmatrix} x_1^{(\xi,1)} & x_2^{(\xi,1)} & \cdots & x_{N_x}^{(\xi,1)} \\ x_1^{(\xi,2)} & x_2^{(\xi,2)} & \cdots & x_{N_x}^{(\xi,2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(\xi,N_a^{(\xi)})} & x_2^{(\xi,N_a^{(\xi)})} & \cdots & x_{N_x}^{(\xi,N_a^{(\xi)})} \end{pmatrix} \tag{2.9}$$

Where $x_n^{(\xi,i)}$ illustrates the n^{th} features of the i^{th} atom in the compound ξ.

This matrix results in a high dimensional representation. In order to produce a descriptor of N_x dimensional space, a transformation is introduced by applying representative quantities

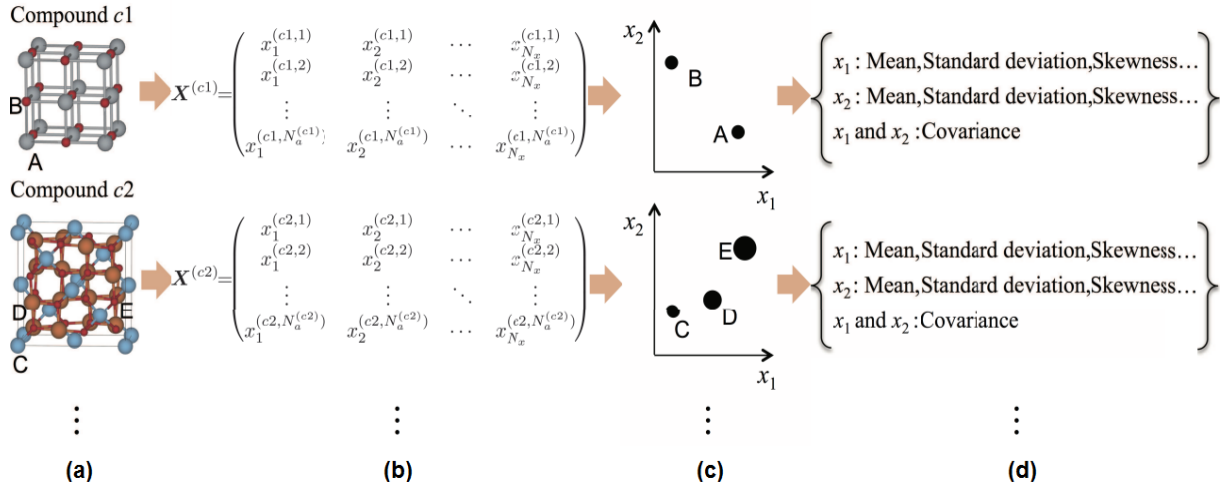


Figure 2.2: Schematic overview of compound descriptor generation. (a) Compounds, (b) matrix representation, (c) data points distribution, (d) representative quantities transforming the distribution into descriptors [58].

such as skewness, covariance, standard deviation, mean, and kurtosis of the distribution as illustrated in Figure 2.2.

The atomic representations of the $X^{(\xi)}$ matrix defined above are a set of 22 elemental features combined with a structural representation chosen from PRDF (previously discussed), GRDF (generalized radial distribution function), BOP (bond-orientational order parameter), and AFS (angular Fourier series).

The study [58] was validated by implementing KRR, Gaussian process regression, and Bayesian optimization (BO) to predict the cohesive energy, lattice thermal conductivity (LTC), and melting temperature using the aforementioned descriptors for input data gathered from DFT computations and experiments.

2.3.1.4 Topological-based descriptor

A. Fedorov et al. came up with a non-conventional way to describe crystal structure in [59]. They first represent a crystal structure through a graph where nodes correspond to atoms and edges to bonds. Figure 2.3 is an illustration of such a graph for NaCl.

This topology representation results in a 2D graph for which each node has a set of features to be provided as input for the ML model. To describe nodes, different features were used to whether predict the molar heat capacities and standard molar entropy or lattice energy. Data was gathered from databases including ICSD and COD (Crystallography Open Database)

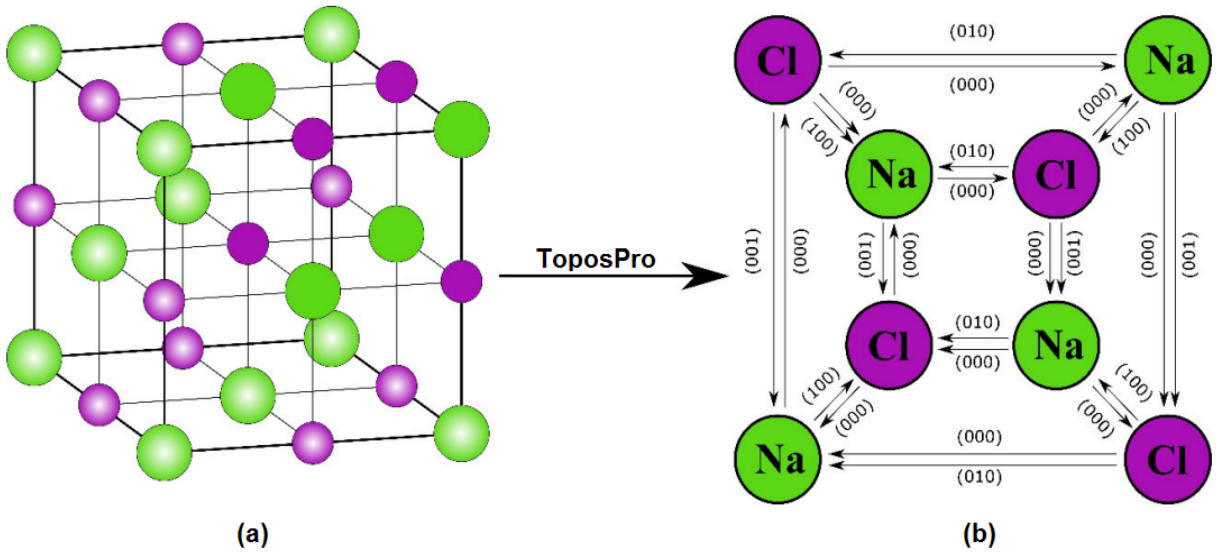


Figure 2.3: Representation of (a) NaCl through the (b) proposed graph [59].

[60], on which the proposed descriptor was applied. To validate the study, the resulting data representation was introduced as an input for an ANN (Artificial Neural Network) model with an architecture of two hidden layers and BFGS (Broyden-Fletcher-Goldfarb-Shanno) learning algorithm.

2.3.1.5 Property-Labelled Materials Fragments (PLMF)

PLMF [61] is a universal crystal structure representation constructed by first determining the atomic connectivity of the material. Then, through a computational geometry approach, the neighbor search is performed and an infinite periodic graph with property labelling is constructed. To better illustrate this process, Figure 2.4 shows the steps of PLMF construction and its constitution.

The adjacency matrix A is represented through Figure 2.4 (c), its entries are defined as follows:

$$x_{ij} = \begin{cases} 1 & for \quad i \qquad connected - to - j \\ 0 & for \quad i \quad not - connected - to - j \end{cases} \tag{2.10}$$

By multiplying this matrix by the reciprocal matrix of square distance D defined as:

$$x_{ij} = \frac{1}{r_{ij}^2} \tag{2.11}$$

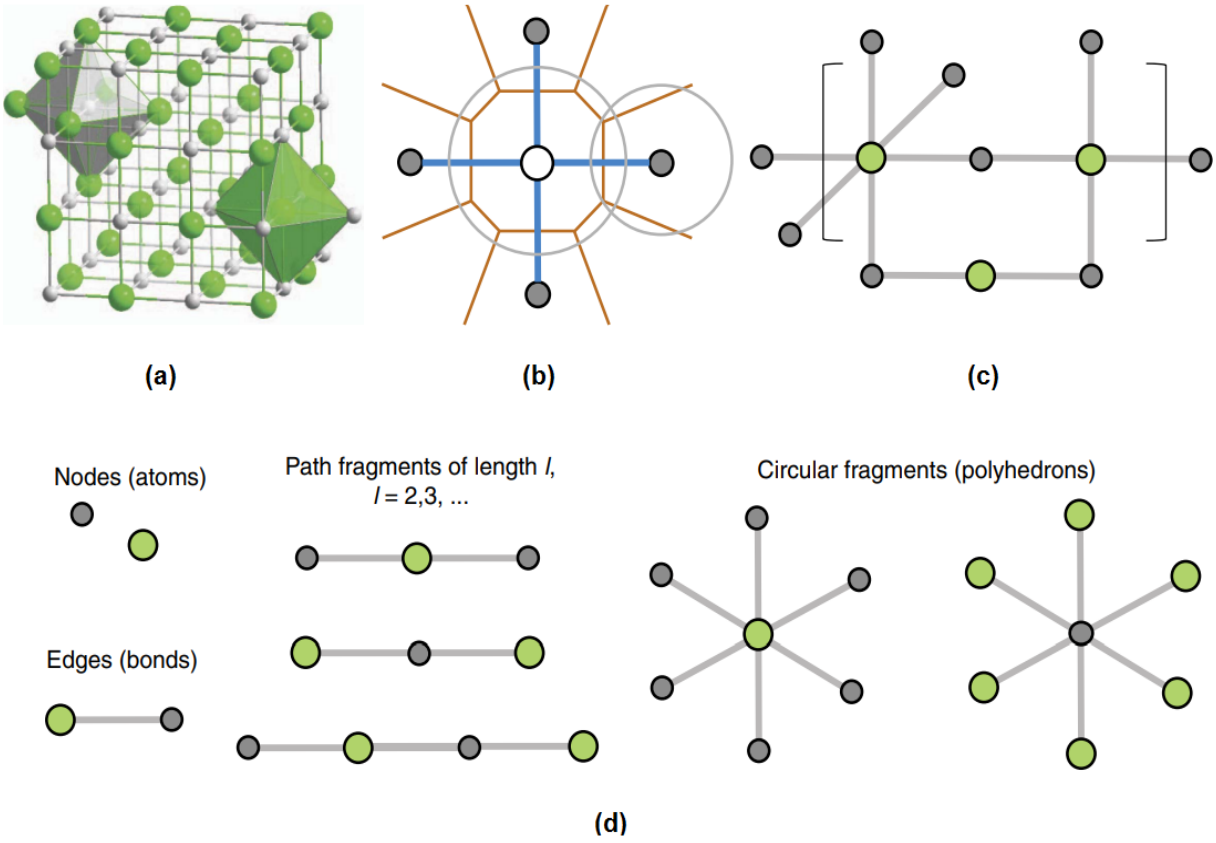


Figure 2.4: PLMF crystal structure descriptor schema. (a) Input crystal structure, (b) neighbors search, (c) infinite periodic graph construction and property labelling, (d) decomposition into fragments and simple subgraphs [61].

one will obtain an $n \times n$ matrix $M = A \cdot D$ with n being number of atoms in the unit cell. From this matrix, the descriptors of the reference property denoted as q can be calculated by the Equations 2.12 and 2.13 for all pairs of atoms (i, j) and only pairs of bonded atoms (i, j), respectively.

$$T^E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |q_i - q_j| \, M_{ij} \tag{2.12}$$

$$T^E_{bond} = \sum_{\{i,j\} \in bonds} |q_i - q_j| \, M_{ij} \tag{2.13}$$

The reference properties include several general properties, measured properties, and other lattice parameters. Authors in [61] proposed and used this descriptor to represent materials gathered from the AFLOW (automatic flow for materials discovery) database [62] and classify them as metal/insulator, and to predict properties such as band gap energy, bulk/shear moduli, Debye temperature, heat capacities. For the modeling stage GBDT (gradient boosting decision tree) technique was performed.

2.3.1.6 2D diffraction fingerprint

It is well known that CNN (Convolutional Neural Networks) ML model is very powerful for image classification [63]. A. Ziletti et al. [64] relied on this model for crystal system classification. For this purpose, they have developed an image-like crystal structure descriptor. This 2D matrix representation is generated by simulating X-ray radiation on the data at hand resulting in a simulated XRD pattern as illustrated in Figure 2.5.

2.3.1.7 Machine learning interatomic potentials (MLIP)-based descriptors

This type of descriptors is known as atom-wise representation instead of structure-wise one. MLIP-based descriptors are very accurate and precise since they include relevant information about each atom neighboring. They are most suitable for energy prediction as they consider atom energy contributions to the total energy [65, 66]. Nevertheless, they can be generalized as universal descriptors for crystal structures with a variety of properties [67].

Table 2.1 summarizes the examined approaches of crystal structure representation for machine learning purpose.

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Faber [55] 2015	Comput.	MP	Formation energy	KRR	Ewald sum-based CM, Extended CM, Sine matrix
Schutt [50] 2014	Comput., experim.	Generated, ICSD, PbAl, CBN	DOS	KRR	PRDF
Seko [58] 2017	Comput., experim.	Generated	Cohesive energy, LTC, melting temperature	KRR, Gaussian process regression, BO	Elemental and structural descriptors
Fedorov [59] 2017	Experim.	ICSD, COD	Molar heat capacities, standard molar entropy, lattice energy	ANN	Topological descriptor
Isayev [61] 2017	Comput.	AFLOW	Metal/insulator, band gap energy, bulk/shear moduli, Debye temperature, heat capacities	GBDT	PLMF
Ziletti [64] 2018	Comput.	AFLOW	Crystal system	CNN	2D diffraction fingerprint

Table 2.1: Summary of crystal structure representation approaches for the use of machine learning.

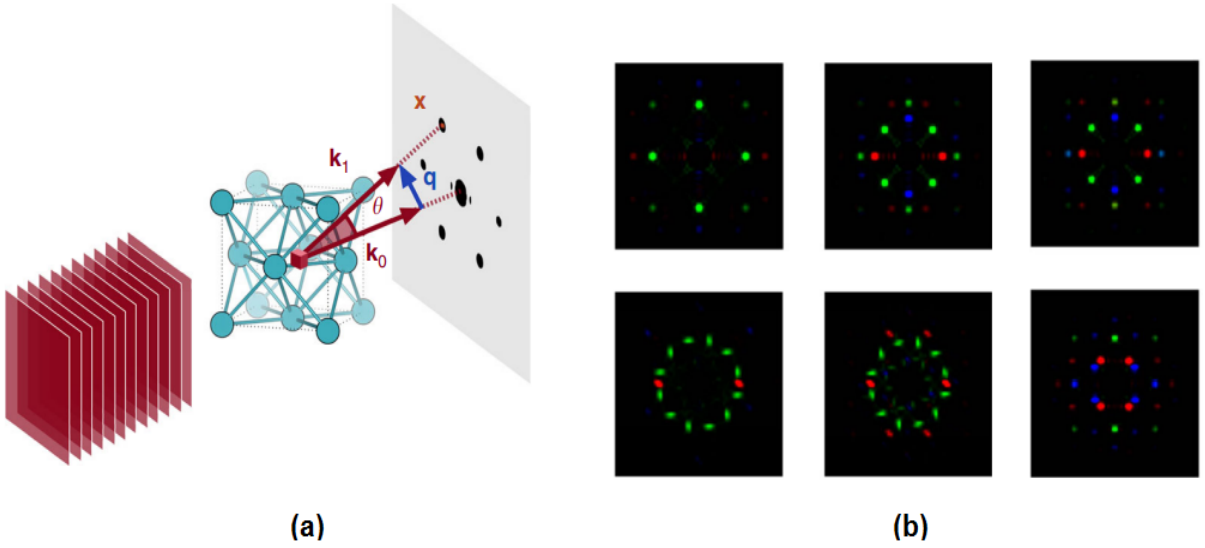


Figure 2.5: X-ray radiation simulation on crystallographic data. (a) 2D diffraction fingerprint computation, (b) resulting image-like 2D diffraction patterns [64].

2.3.2 Crystal structure prediction approaches with machine learning

In materials science, one of the most challenging problems is crystal structure prediction [68], whether it is about the discovery of new materials or the prediction of crystal structures' properties. Lately, machine learning has become a key solution to laborious and time consuming problems. Its applications are found in any imaginable field. In the research field of the present study, ML methods and algorithms have been widely employed. The subsections below represent a review of current studies related to crystal structure prediction.

2.3.2.1 Decision tree-based approaches

As their name suggests, decision trees are used to formally and graphically reflect decisions and decision making using a decision-tree-like approach. They include several models in machine learning. A decision tree is composed of a root node representing the condition, intermediate nodes which are children-node-alternatives for their father-node, and leaf nodes of the decision.

W. Tong et al. [69] conducted a machine learning-based study for the prediction of the elastic modulus property. For this purpose, authors first examined different machine learning models to predict other properties in order to determine the best performing model. Random

Forest (RF), Support Vector Machine (SVM) as a regressor, and a Deep Neural Network (DNN) were used for the prediction of bulk modulus, Young's modulus, and shear modulus properties. The main aim for this study is to accelerate properties prediction in the materials search process. As depicted in Figure 2.6, after the structure generation computation that was performed using CALYPSO code [70], and once the structures are optimized, the selected ML model is used to replace the time consuming DFT-based calculation for properties prediction.

Carbon materials were extracted from the SACADA (Samara Carbon Allotrope Database) database [71] and represented through 15 MATMINER-based [72] descriptors. random forest algorithm is reported to have given the best performance among other models and was selected for the prediction of the elastic modulus property. The full execution of this process resulted in the discovery of a brand-new carbon phase which has never been reported before.

M. Amsler et al. [73] have also used random forest and decision trees to predict band gap energy and formation enthalpy. They selected ternary compound from the computational Open Quantum Materials Database (OQMD) and used Magpie (Materials-Agnostic Platform for Informatics and Exploration)-based features to represent ML inputs. For the modeling stage, data was first divided into subsets of similar materials and each subset was trained separately using decision trees to predict the band gap energy. In addition, random forest algorithm was used for the prediction of the formation enthalpy.

Decision trees-based approaches can also be used for classification problems. Several studies on crystal structure prediction in the literature relied on decision trees classifiers. In [74], binary compound data was collected from OQMD labelled with the crystal system. For ML purpose, the data was represented by a set of 8 features including the number of atoms of type A, the number of atoms of type B, atomic numbers of A and B, electro-negativities of A and B, and atomic radius of A and B. First, an unsupervised machine learning model based on Gaussian mixture was applied to expose material data structure. As a result, two data clusters were revealed; the first cluster contains data fitting into the predefined 492 prototype structures, and the second represents the data that does not fit in. Then, random forest algorithm with a 100 trees was trained in a supervised manner on the first cluster of data, where each material has been labelled with one of the 492 prototypes. The label was used to classify this data according to the assigned prototype structure. The average cross-validation accuracy achieved for the multi-class classification in this work is of 79%.

Authors in [75] conducted a study on perovskites materials generated computationally. The data is labelled according to the crystal system as the target value. It is to be mentioned that only four crystal systems were considered, namely cubic, orthorhombic, tetragonal, and

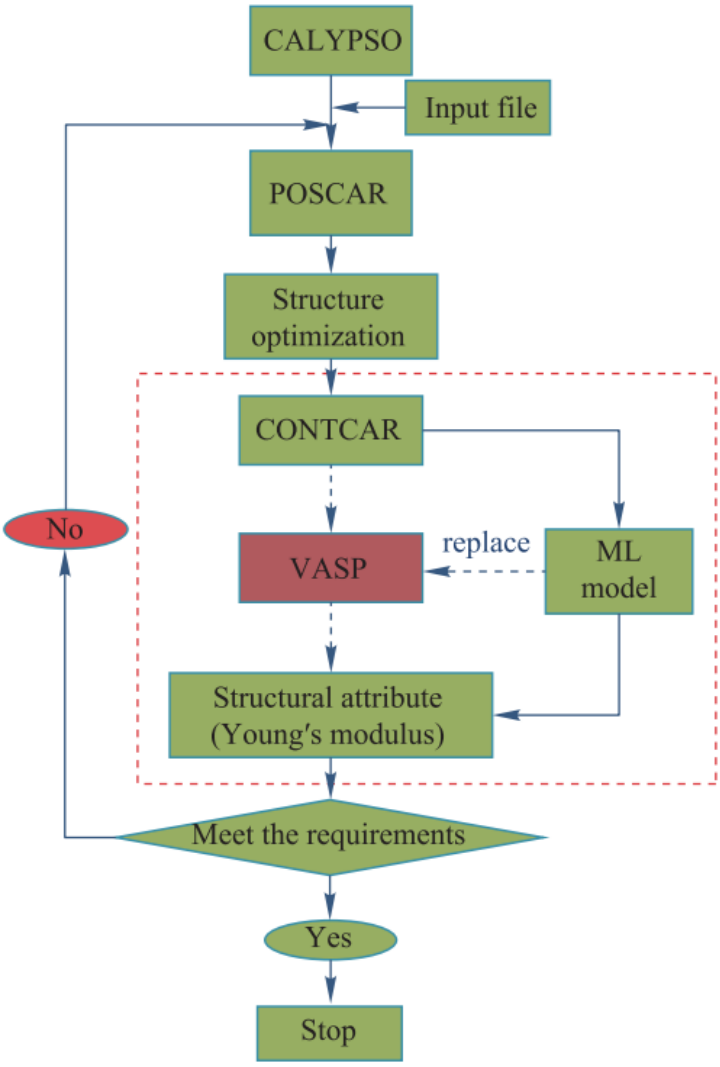


Figure 2.6: The proposed flow chart for an accelerated prediction of crystal structures using machine learning [69].

rhombohedral, the rest of the crystal systems were omitted since they do not satisfy the perovskite structural criterion. A set of 13 features were used as descriptors to represent the examined data. For this multi-class classification task, Light GBM (Gradient Boosting Machine) algorithm was selected. It is a decision tree-based algorithm which combines a series of individual trees to form a strong learner where each new tree's mission is to minimize the previous trees' error. Moreover, rather than growing vertically (level-wise) like other tree-based models, Light GBM grows horizontally (leaf-wise) where the leaf that is selected to grow on is the one with the greatest delta loss. The classification accuracy achieved by Light GBM on the generated perovskites is 80.3%.

Random forest in particular had proved to be quite successful in terms of multi-class classification in crystal structure prediction, especially in terms of crystal system and space group classification. Authors in [76, 77] have extracted data from the MP database labelled with the crystal system and the space group. In an effort to define enhanced descriptors to represent the data, Y. Li et al. [76] proposed composition-based features combined with Magpie. For the modeling stage, random forest, XGBoost (Extreme Gradient Boosting), and deep neural networks were used for learning and classifying data according to the crystal system and space group. The three models' hyperparameters and architecture were defined as follows:

- entropy was selected as criterion, number of trees set to a 100, number of features to 80, the max depth to None, and finally, 2 and 1 as min samples split and min samples leaf, respectively.

- XGBoost: the booster was selected as "gbtree" with 6 as max depth, and the number of trees was 180. Alpha, gamma, lambda, and learning rate were defined as 0, 0, 1, and 0.3, respectively.

- DNN: the architecture of the network was chosen to have 7 fully connected layers all activated with ReLU (Rectified Linear Unit) in addition to Dropout and BatchNorm after every layer except the last one with the purpose to avoid overfitting. Cross entropy and Adam (adaptive moment estimation) optimizer were used as the loss function and optimization algorithm, respectively, while the hyperparameters were tuned by setting the number of epochs, the batch size, and the learning rate to 2000, 255, and $10e^{-2}$, respectively.

The results of this study show that the random forest model yielded the best results with a score of (0.835 – 0.829) for the metrics (Accuracy – F1-score).

Y. Zhao et al. [77] worked on the same classification task of space group and crystal system. They opted for a Multi-Layer Perceptron (MLP) and random forest to map the MP gathered data representation to the output classes. Three different features descriptors were selected, namely Magpie, atom vector, and atom frequency. The three descriptors were combined with the two learning algorithms in order to select the best descriptor-algorithm combination for the classification of crystal system and space group. The ML models were used as follows:

- MLP: two architectures were selected depending on whether the classification is one-versus-all-based or multilabel-based. The first architecture has a total of 11 layers while the second one is composed of 13 ones. ReLU is used as an activation function for all layers except the last ones in each architecture. The last layers were activated using sigmoid or softmax.
- RF: the number of decision trees was defined as 50 where each was trained on a subset with sample features that are randomly selected.

As a result, the random forest model combined with Magpie outperformed all other combinations in terms of crystal system and space group classification with a score of (0.650 – 0.591) and (0.765 – 0.566), respectively, for the metrics (F1-score – MCC).

Another type of input to consider is XRD patterns. In [78, 79], authors have used such inputs simulated from the ICSD database. Y. Suzuki et al. [78] chose to represent this data using ten peaks-based information as well as diffraction peaks number. This descriptor was fed to a random forest model for the task of XRD patterns classification into their respective crystal system and space group. The implemented RF model was able to perform this classification with an accuracy of 93.07% and 83.62% for crystal system and space group prediction, respectively.

In [79], a set of five ML models including logistic regression (LR), K-nearest neighbor (KNN), DT, RF, and extremely randomized trees (ExRT) were implemented to perform the same classification task on the same type of data. An eleven features set based on peaks information was used to describe the XRD patterns and fed to the previously mentioned algorithms tuned by random search. The ML models trained and tested on the describe data were compared in terms of crystal system and space group classification. ExRT, the RF-based ML model yielded the best performance; the edge it took over the other models is due to its randomly chosen decision-making variables which significantly reduce overfitting. The accuracy achieved for crystal system and space group classification is 90% (except for triclinic system being rare in ICSD) and 88%, respectively.

The summary of the examined studies related to DT-based crystal structure prediction is presented in Table 2.2.

2.3.2.2 Support vector machine-based approaches

SVM learning algorithm is considered as a powerful classification/regression model. It is favored for its significant accuracy with a low computational power. Its applications cover many fields including that of materials science. As an example of SVM classification in crystal structure prediction, authors in [80] have proposed expandable features to describe alloy materials. This descriptor is generated by transforming $\{n_d^{(N)}, \sigma_d^{(N)}\}$ (orbital occupancy, orbital spin) to $\{n_d^{ex}, \sigma_d^{ex}\}$ using regression tree ensembles (Figure 2.7 (b)). The dataset is composed of binary and ternary alloys as well as high entropy alloys (HEA), where only binary alloys were considered for the training process (Figure 2.7 (a)). The dataset was generated using DFT-based Akai-KKR-CPA (Akai-coherent potential approximation to korringa-kohn-rostoker) code. In the modeling stage, SVM algorithm with ECOC (error-correcting output coding) and Gaussian kernel function were used to classify data according to its structural phase (Figure 2.7 (c)).

The resulting performance of the implemented SVM achieved an accuracy of 80.56% for structural phase classification of alloys and 84.20% for that of HEA.

When applied in regression tasks, SVM is usually referred to SVR (Support Vector Regressor). S. Jarin et al. [81] used SVR-based approach among other models to predict the lattice parameters of perovskite materials. For this purpose, authors gathered a total of 2225 experimental and theoretical ABO_3 from [82] and represented them with a set of 12 atom-based features. Data was first classified according to its crystal system using RF, SVM genetic algorithm (GA)-SVM, NN, GA-NN. These ML models were tuned as follows:

- RF: number of trees with 200 estimators, minimum sample split and maximum depth set to 2 and 29, respectively.
- SVM: C and γ parameters were set through grid search to 0.853 and 0.003, respectively.
- NN: Levenberg–Marquardt training algorithm was used for the backpropagation process and sigmoid for transfer function.

GA-NN model was able to outperform other models in crystal system classification with an accuracy score of 88%.

Once the crystal system predicted, authors proceeded to lattice parameters prediction using SVR and GA-SVR with RBF (radial basis function) kernel. Both models yielded good

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Tong [69] 2020	Comput.	SACADA	Bulk modulus, Young's modulus, shear modulus	RF, SVM, DNN	MATMINER-based
Amsler [73] 2019	Comput.	OQMD	Band gap energy, formation enthalpy	RF, DT	Magpie-based
Takahashi [74] 2019	Comput.	OQMD	Prototype structure	Gaussian mixture, RF	Defined
Behara [75] 2021	Comput.	Generated perovskites	Crystal system	Light GBM	Defined
Li [76] 2021	Comput.	MP	Crystal system, space group	RF, XGBoost, DNN	Composition-based combined with Magpie
Zhao [77] 2021	Comput.	MP	Crystal system, space group	MLP, RF	Magpie, atom vector, atom frequency
Suzuki [78] 2018	XRD	ICSD	Crystal system, space group	RF	Defined
Suzuki [79] 2020	XRD	ICSD	Crystal system, space group	LR, KNN, DT, RF, ExRT	Defined

Table 2.2: Summary of DT-based crystal structure prediction approaches.

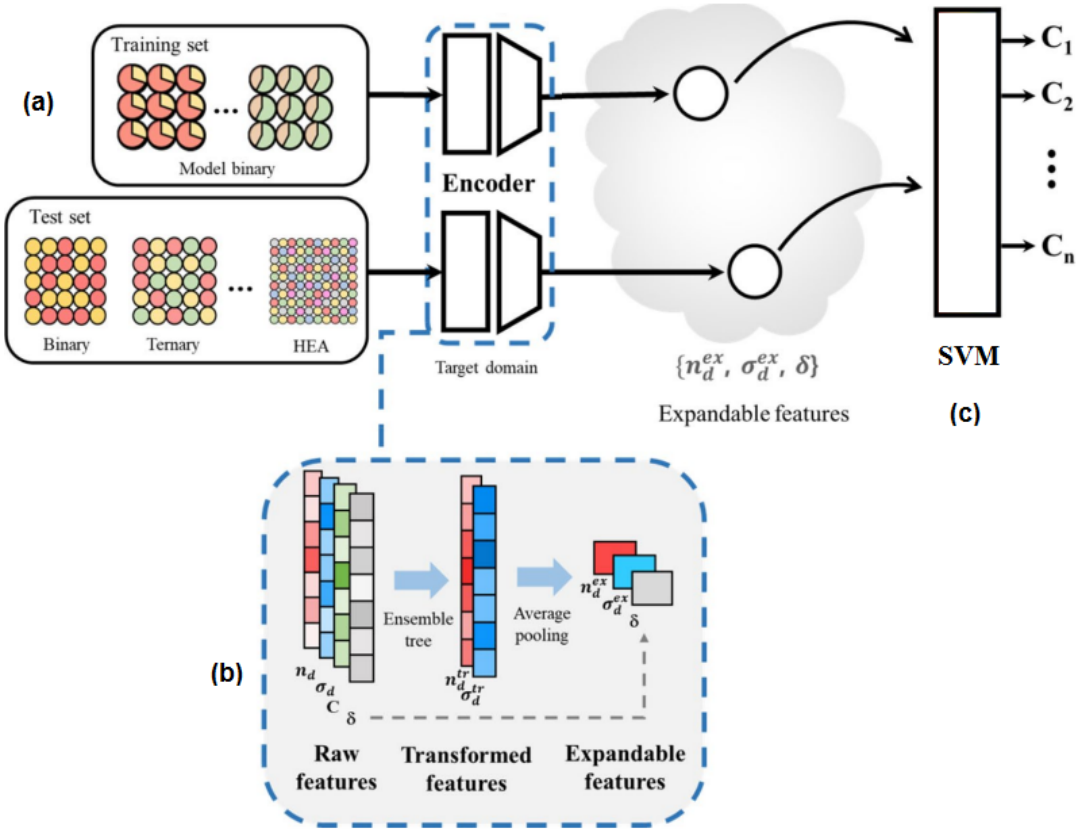


Figure 2.7: Flowchart of the expandable features generation and structural phase classifier. (a) Training/testing sets, (b) raw features into expandable features transformation, (c) SVM classifier [80].

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Jin [80] 2021	Comput.	Generated alloys	Structural phase	SVM	Defined expandable features
Jarin [81] 2022	Comput., experim.	ABO_3 [82]	Crystal system, lattice parameters	RF, SVM, GA-SVM, NN, GA-NN	Defined

Table 2.3: Summary of SVM-based crystal structure prediction approaches.

prediction results for the lattice parameters (a, b, and c) with an accuracy of 95% for the GA-SVR model.

Summarized details of SVM-based crystal structure prediction approaches are presented in Table 2.3.

2.3.2.3 Neural network-based approaches

Neural networks are ML models for which the structure and nomenclature are modeled after the human brain. They reflect and mirror the communication and signalization between biological neurons. NN models are very powerful and form the core of deep learning algorithms. They have been used extensively for crystal structure prediction problems.

M. Kusaba et al. proposed in [83] a metric learning framework using ML models including MLP. For a given chemical composition, the proposed framework is able to automatically choose template structures by element substitution for the unknown stable structure. Authors first gathered data from the MP database and constructed a labelled dataset where an instance input corresponds to a pair of chemical compositions and the output to whether this pair is identical or not. Two chemical compositions are considered identical if they're similar to a certain extent. Among the applied models (Siamese network, keep it simple and straightforward-KISS, MLP classifier, and MLP regressor) to perform this binary classification; the binary classification based on MLP yielded the best performance. The descriptor used to represent the chemical compositions is a XenonPy-based 58×5 (290-dimensional) vector. The result this framework achieved is a score of 0.991, 96.4%, 96.3%, and 96.6% for the metrics AUC (area under ROC curve), accuracy, sensitivity, and specificity, respectively.

In the work cited in [68], a deep neural-network model was implemented to compare the atomic sites topologies in known crystal structures and use the information learned from such comparison to forecast potential compositions of unidentified substances that a synthetic

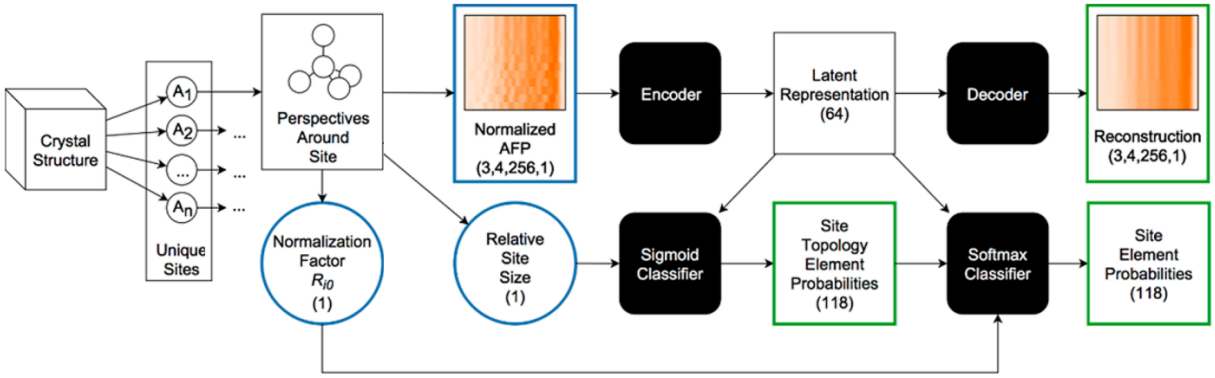


Figure 2.8: Proposed approach scheme for chemical elements classification [68].

chemist may explore. For this purpose, experimental data was gathered from the ICSD and COD databases. Data was represented through atomic fingerprints (AFP) inspired from CSFP (crystal structure fingerprint) [68] and defined as follows:

$$AFP_i^k(R) = \sum_j \delta(R - \frac{R_{ij}^k}{R_{i0}}) \tag{2.14}$$

Where i represents the atom at perspective k, δ denotes a delta function, R_{ij} and R_{i0} are the distance between atoms i and j and the distance between i and i's nearest neighbor, respectively.

Before feeding this input to the DNN classifier several steps were performed as highlighted in Figure 2.8.

The extracted AFP goes through a VAE (variational autoencoder) in order to allow the DNN to learn a simplified 64-dimensional representation of AFPs. In addition to this generated representation, normalized geometric descriptors including non-normalized R_{i0}, and the crystal structure's smallest interatomic ration to R_{i0} are fed to a sigmoid classifier with 5 layers. The number of the classifier's outputs corresponds to the number of chemical elements in the periodic table (118). The resulting output combined with the non-normalized geometric descriptors are then introduced as input to a softmax classifier with five layers and the same number of outputs as the former one.

Adam optimizer was used to train the DNN where the learning rate was initially set to 3×10^{-7} then it was increased to 3×10^{-5} by a step of 3×10^{-10} every batch. Dropout layers were added with a 0.05 probability with regularization $L2 = 5 \times 10^{-4}$ and a batch size of 16. The obtained results show that this scheme is able to predict chemical elements with an error rate of 31%.

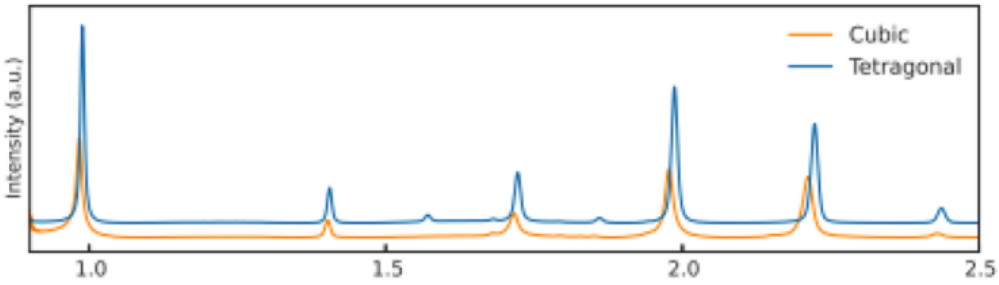


Figure 2.9: Example of two XRD patterns of two different crystal systems (cubic in orange and tetragonal in blue) [84].

Another classification problem was examined in [45, 84]; similar to previously discussed studies, authors in [45, 84] proposed a CNN-based approache to classify crystal structures. The input data represents XRD patterns extracted from ICSD. Figure 2.9 depicts two different crystal system XRD patterns.

In [45], authors attempted to classify crystal structures according to their crystal system, extinction group, and space group. Unlike the majority of XRD-based crystal structure prediction studies, this work considers the XRD pattern raw images as the final input of the ML model without a features engineering process. Three CNN models were developed to classify crystal structures into their respective crystal system, extinction group, and space group. The common architecture between the three CNNs is composed of an input layer, three convolutional layers followed each by an average pooling layer, and the flattened resulting multi-dimensional vector. Then each CNN architecture is followed by two fully connected layers and an output layer where the number of nodes differ from one architecture to another considering the fact that the number of classes is different from one case to another. ReLU is used as an activation function and dropout layers were added with a 30% probability. The classification accuracy achieved 81.14%, 83.83%, and 94.99% for crystal system, extinction group, and space group, respectively.

A. Chakraborty et al. in [84] conducted a study for crystal system classification. The data was converted into a vector to be considered as an input for the learning process. Several ML models namely NB (Naïve Bayes), KNN, LR, RF, SVM, GBRT (gradient boosted regression trees), and MLP were implemented for comparison with the proposed CNN-based approach. The designed all-Convolutional Neural Network, referred to as a-CNN is distinguished from other CNNs by the fact that its architecture does not include max-pooling layers between convolutional ones. The crystal system classification performance achieved by a-CNN exceeds

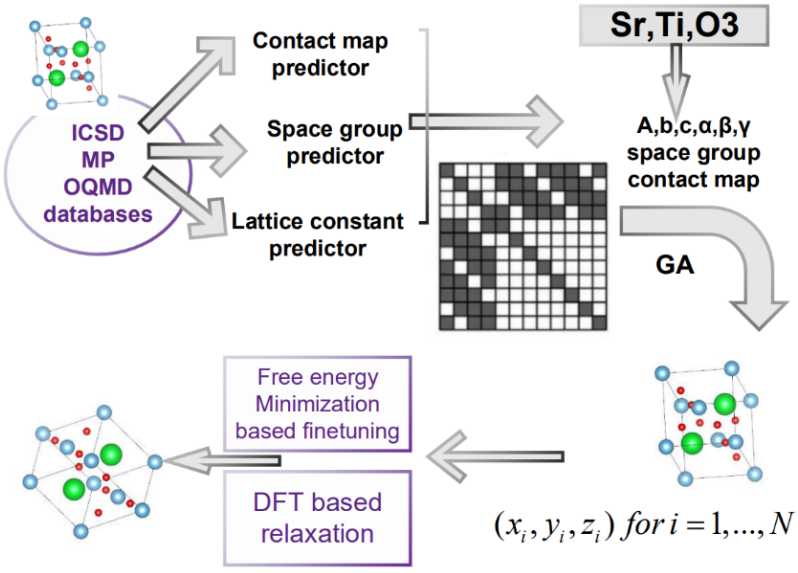


Figure 2.10: AlphaCrystal's flow chart for crystal structure reconstruction using residual neural networks [85].

that of all other implemented ML models with an accuracy and F1-score of 95.6% and 0.949, respectively.

J. Hu et al. in [85] have used another type of deep neural networks to resolve a classification problem related to crystal structure prediction. First, computational data was gathered from the MP database, and represented through a defined 11D-element-wise features descriptor. Then, residual neural networks are used to predict the contact map, space group, along with lattice constants. The proposed framework named AlphaCrystal is able to learn geometric patterns and the distribution of atom interactions and use this hidden knowledge to predict the contact map. Once the contact map predicted, the 3D target crystal structure can be reconstructed using genetic algorithms. For the validation process, the predicted structures were relaxed using the DFT-based VASP (Vienna ab initio simulation package) tool. Figure 2.10 illustrates the general scheme of this approach.

According to the learnt knowledge about atomic interaction distribution, the contact map (matrix) can be defined as follows:

$$x_{ij} = \begin{cases} 1, & if R_{ij} \in [a+b-0.4, a+b+0.4] \\ 0, & else \end{cases} \tag{2.15}$$

Where i and j are the pair of atoms, R_{ij} the distance between them, and a and b the covalent radius of i and j, respectively. Figure 2.11 shows a predicted contact map matrix

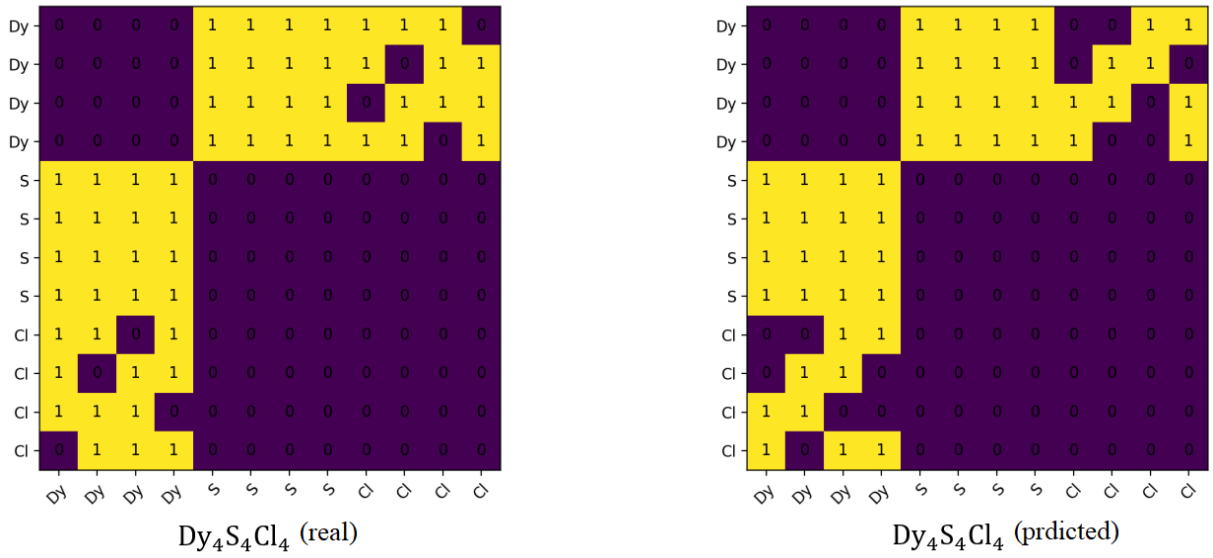


Figure 2.11: Example of a contact map prediction (real Vs. predicted) [85].

compared to a real one.

ReLU activation function was used in the residual neural network with Adam optimizer and the cross-entropy loss function since it's adapted for a binary output. The number of epochs was set to 125 and the learning rate to 10^{-3}. The obtained results achieved an average accuracy of 0.8543 and an average score of 0.2407 and 0.193 for the metrics RMSD (root mean square distance) and MAE (mean absolute error), respectively.

Table 2.4 summarizes the examined NN-based crystal structure prediction approaches.

2.3.2.4 Graph network-based approaches

GNNs, short for Graph Neural Networks, are a family of deep learning algorithms that were created to perform inference on graph-represented data. It is composed of nodes and edges representing the data features and the relationship between the data points [86].

GNNs are defined as neural networks that may be used to analyze graphs directly; they're found in applications where data instances are related to each other such as social networking. As powerful as CNNs can be, they fail at what GNNs can do. The reason is that CNNs cannot be directly performed on graphs because of their complex topology and size. In addition, the order of nodes varies and, unlike GNNs, CNNs cannot deal with unfixed ordering.

Since crystal structures are composed of atoms bonded together through different types of bonds, GNNs might be adequate candidates to represent them. In [87] T. Xie et al.

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Kusaba [83] 2022	Comput.	MP	Similarity (identity)	MLP classifier, MLP regressor Siamese network, KISS	XenonPy-based features
Ryan [68] 2018	Experim.	ICSD, COD	Chemical elements	DNN	AFP
Park [45] 2017	XRD	ICSD	Crystal system, extinction group, space group	CNN	XRD patterns
Chakraborty [84] 2022	XRD	ICSD	Crystal system	a-CNN, NB, RF, KNN, DT, SVM, GBRT, MLP	Defined
Hu [85] 2021	Comput.	MP	Contact map, space group, lattice constants	Residual NN	Defined

Table 2.4: Recapitulation of NN-based crystal structure prediction approaches.

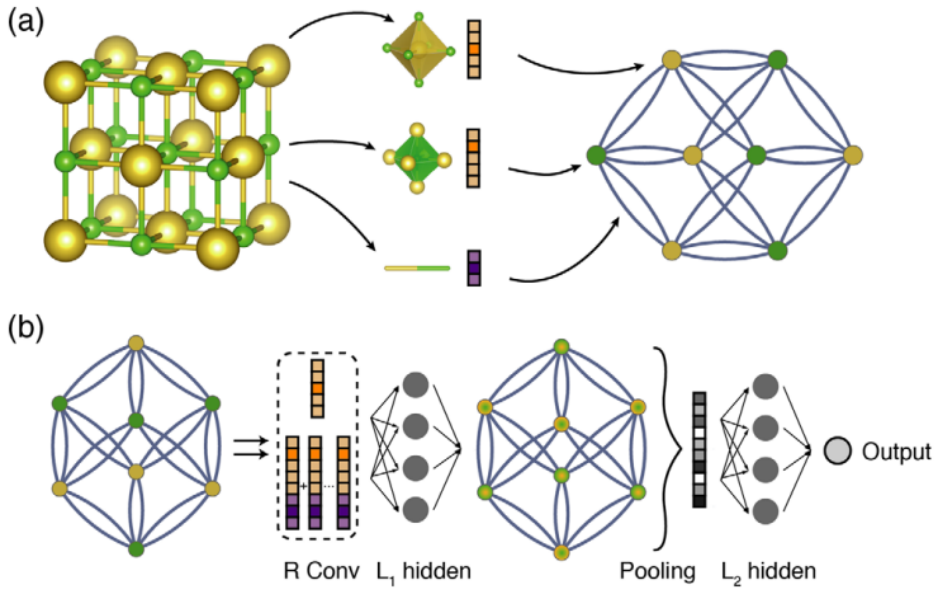


Figure 2.12: Depiction of the crystal graph convolutional neural network. (a) Crystal graph construction, (b) convolutional neural network built on top of the graph [87].

proposed a GNN-based approach for crystal structure property prediction named CGCNN (Crystal Graph Convolutional Neural Network). The general scheme of their approach is illustrated in Figure 2.12.

In the constructed crystal graph, nodes correspond to atoms and edges to bonds. Nodes and edges are both represented through a vector. Convolutional hidden layers are applied on top of this constructed graph resulting in another graph. Then, pooling and other hidden layers are applied to the second graph in order to predict the output. To validate this framework, data was gathered from the MP database for the prediction of seven different properties. The MAE score achieved for formation energy, absolute energy, Fermi energy, band gap, bulk moduli, shear moduli, and Poisson ratio prediction is of 0.039, 0.072, 0.388, 0.363, 0.054, 0.087, and 0.03, respectively.

An extension of this work was conducted by S. Louis et al in [88]. They used the same crystal-to-graph transformation and examined the same set of properties to predict, while the architecture of the GNN was based on graph attention (GAT). GAT networks however present limitations with regard to edge information in contrast with neighboring nodes information which is well characterized. Consequently, authors in this study augmented GAT layers with connecting edges information; these layers are referred to as AGAT. In addition, a global attention layer is proposed in the architecture added after the AGAT layer. This new layer is

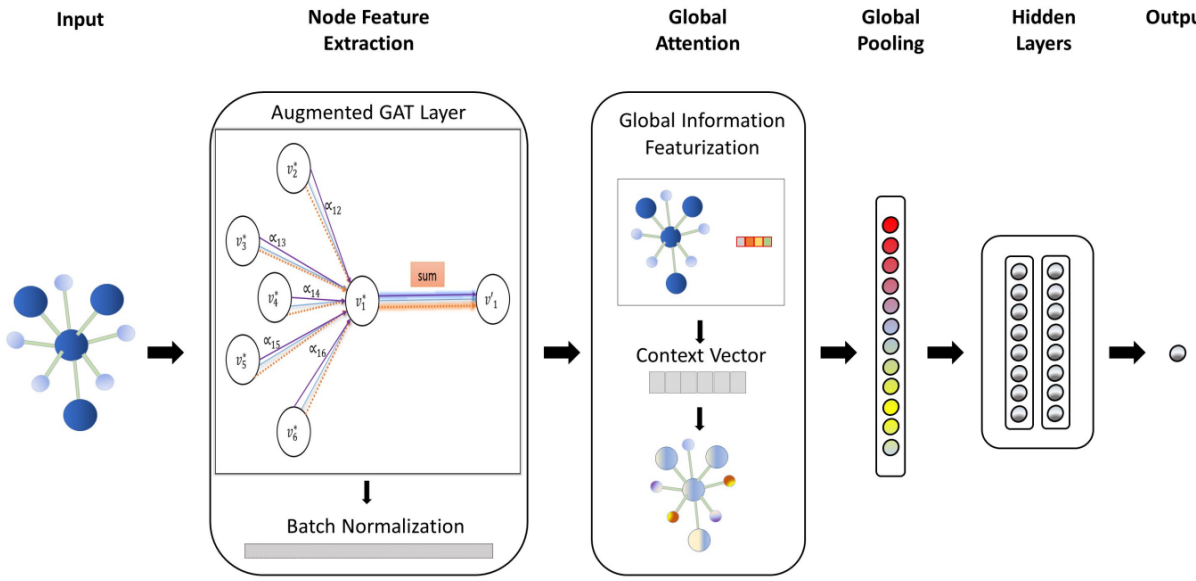


Figure 2.13: Architecture scheme of the proposed global attention graph CNN model (GAT-GNN) [88].

added with the purpose to translate the locally learned information from node and edge level to graph level for better interpretability. Figure 2.13 presents the proposed GNN architecture.

The output of the regression is predicted by applying a global pooling layer on global attention one, then, some hidden layers to the global pooled resulting layer. The model was trained for 500 epochs at most, the loss function and optimization algorithm were set to Smooth L1loss and Adam. The learning rate was initiated at 5×10^{-3}, then it was decreased to 5×10^{-4}, and then to 5×10^{-5}, and finally, the batch size was set to 256. The results obtained achieved an MAE score of 0.039, 0.048, 0.33, 0.322, 0.047, 0.085, and 0.029 for the properties formation energy, absolute energy, Fermi energy, band gap, bulk moduli, shear moduli, and Poisson ratio, respectively.

Crystal Graph-based models can predict a wide range of properties with high accuracy. Although they encode interatomic interactions, they ignore traits that include features of orbital-orbital interaction. For this particular reason, authors in [89] proposed Orbital Graph Convolutional Neural Network, referred to as OGCNN, which involves orbital-orbital interactions. For this purpose, orbital-field matrix (OFM) was used to represent data generated computationally using DFT calculations. In order to predict the formation energy, band gap, and Fermi energy, OGCNN was trained for 100 epochs with an MSE (mean squared error) loss function and SGD (stochastic gradient descent) optimization algorithm. OGCNN

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Xie [87] 2018	Experim.	ICSD	Formation energy, absolute energy, Fermi energy, band gap, bulk moduli, shear moduli, Poisson ratio	GNN-based CGCNN	Defined
Louis [88] 2020	Comput.	MP	Formation energy, absolute energy, Fermi energy, band gap, bulk moduli, shear moduli, Poisson ratio	GNN-based GATGNN	Defined
Karamad [89] 2020	Comput.	Generated	Formation energy, band gap, Fermi energy	GNN-based OGCNN	OFM
Cheng [90] 2022	Comput., experim.	OQDM, MatB	Formation enthalpy	GNN-based	Defined

Table 2.5: Summary of graph network-based crystal structure prediction approaches.

achieved an average MAE of 0.0466 for the formation energy prediction and 0.32 and 0.38 for band gap and Fermi energy, respectively.

Another graph network-based approach for crystal structure prediction was proposed in [90]; it is a framework in the form of (database + GN + optimization algorithm). The databases separately used are OQMD and MatB (Matbench) [91] with a graph network, while the selected optimization algorithms are random searching (RAS), particle swarm optimization (PSO), and Bayesian optimization (BO). A total of two graph networks were implemented, one for each database, to map the correlation between the input data and the formation enthalpy as the output to predict. Several GN architectures were tested and the ones yielding the smallest prediction error were selected. The results achieved by this framework are MAE values of 0.016 and 0.031 for the OQMD and MatB databases, respectively.

The previously investigated graph network-based approaches are briefly reviewed in Table 2.5.

2.3.2.5 Interatomic potential-based approaches

Interatomic potentials describe the interaction between a pair of atoms or an atom and a group of atoms. They possess both attractive and repulsive components. Interatomic potentials provide the energy of the system as a function of the atomic positions. They are extremely accurate especially in terms of energy, forces, and stress prediction. In PES (potential energy surface) the energy of the system is considered to be the sum of atomic energies. Therefore, accurate and unique atom-wise descriptors are necessary.

H. Wang et al. proposed a deep potential-based approach for crystal structure prediction of (AlMg) binary alloys [92]. For this purpose, authors generated AlMg structures using DFT calculations and described the data using atom-wise local environment information. The modeling stage includes two embedded neural networks; the first one is the embedding NN generating symmetry-reserving descriptors and the second one is the fitting NN mapping the atomic energy to these descriptors. The RMSE (root mean square error) score for the formation energy prediction reached 0.006.

Authors in [93] trained machine learning potentials to predict crystal structures' energies, forces, and stress. They proceeded by computationally generating a set of materials using DFT-based calculation through VASP tool. Data was then represented using Behler-Parrinello's atom centered descriptors [94]. PCA (principal component analysis) was used in ordered to make the input vectors uncorrelated. Neural networks model was used to train this data with Adam optimizer and MSE loss function with L2 regularization term to prevent overfitting. The epoch size is determined in accordance with the MSE value; the training stops when the validation RMSE reaches the value of 0.01 for the energy prediction.

A moment tensor potential (MTP)-based [95] approach was proposed in [48] for crystal structure energy prediction on-the-fly. To this end, computational data was generated and represented through atomic neighborhood-based descriptors. In addition, authors proposed a new concept allowing ML potentials transferability with an active learning approach. The idea behind this concept is to prevent inaccurate predictions when the introduced data to be predicted is farfetched from the model's learning set. The learning stage of MTP uses a regularized linear regression algorithm. The scheme of the proposed approach is illustrated in Figure 2.14.

The input data configuration is first tested to whether it needs to be actively learned or not by checking if a computed parameter named extrapolation grade exceeded the predefined extrapolation threshold or not. If it does, then DFT is used for the prediction and this instance is added to the learning set and the configuration is learned. If not, then energies,

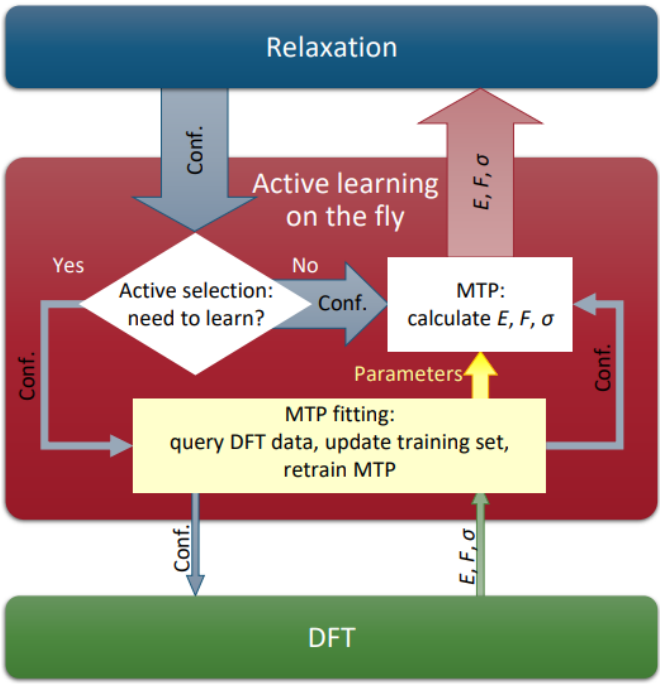


Figure 2.14: Flowchart of the proposed MTP-based approach [48].

forces, and stress are predicted using MPT. This approach was tested with carbon, sodium, and boron allotropes and it yielded an RMSE of 0.011.

To briefly summarize the aforementioned studies, Table 2.6 presents details of the MLIP-based approaches.

2.4 Summary and discussion

As seen in the previous sub-section, many ML-based approaches have been proposed and implemented to solve different crystal structure prediction problems. It is however difficult to fairly compare the reviewed studies because of the diversity of the investigated database, the property to predict, as well as the used performance metrics. Nevertheless, we can divide crystal structure prediction problems into two main families. The first one is classification-based predictions such as the classification of the crystal system, space group, structural phase, etc. In terms of crystal system classification, we notice that all approaches yielded an accuracy above 80%. The studies [78, 79, 45, 84] used the same database (ICSD), where [78, 79] employed RF and [45, 84] CNN-based models. Results achieved by [78] (93.07%) exceed those of [79] (90%) because of the employed descriptors that are more suitable for

Ref.	Data type	Data source	Prediction	ML model	Descriptor
Wang [92] 2020	Comput.	Generated Al-Mg	Energy, force	NN potentials	Atom-wise local environment-based
Hong [93] 2020	Comput.	Generated	Energy, force, stress	NN potentials	Behler-Parrinello's atom centered-based
Podryabinkin [48] 2019	Comput.	Generated	Energy, force, stress	MTP	Defined

Table 2.6: Outline of MLIP-based crystal structure prediction approaches.

crystal system classification. Likewise, [45] achieved 94.99% accuracy compared to [84]'s 95.6% accuracy which reflects the robustness of the CNN modified architecture and the descriptors used with regard to crystal system classification. In addition, the work proposed by [64] was able to achieved the highest crystal system accuracy (100%) also using CNN, thus proving the effectiveness of the 2D diffraction fingerprint descriptors.

The second family is regression-based predictions such as the prediction of energy, bulk moduli, shear moduli, etc. Among all reviewed regression-based studies, the best performing crystal structure representation approaches are [50] and [58] using PRDF and elemental-structural descriptors; they yielded a score of 0.0077 in terms of MAE and 0.0071 in terms of RMSE, respectively. Moreover, the NN potentials-based approach proposed by [92] achieved the best score of 0.006 in terms of RMSE among ML-based crystal structure prediction approaches.

To summarize, we conclude that:

- Deep learning-based approaches through CNN modeling are very powerful with regard to classification tasks especially when the input data is 2D.
- PRDF and elemental-structural descriptors are suitable for regression tasks.
- NN potentials-based approaches are very accurate for the energy prediction (regression task).

2.5 Conclusion

As we conclude the inspection and analysis of the state of the art in CSP, we stand at the crossroads of discovery and innovation. In this chapter, the dynamic environment of CSP research has been revealed, demonstrating the noteworthy advancements, cutting-edge methods, and ongoing challenges that characterize this discipline. Our journey through recent growth, discoveries, and the evolving methodologies has provided us with a comprehensive understanding of the current state of CSP.

With the knowledge acquired from the CSP state-of-the-art, we have drawn conclusions of the potential challenges and opportunities that still arise. In the next couple of chapters, we move forward with our contribution in the crystal structure prediction field, where the data representation and the modeling approaches adopted are an outcome of synthesizing the insights gained in this chapter.

Chapter 3

Crystal Structure Features Engineering

"Torture data and it twill confess to anything." - Ronald Coase.

3.1 Introduction

In machine learning, one of the most essential issues, that is as important as learning, optimization, or generalization, is to define the right input for the learning process. This task is unquestionably important since it directly affects the outcome of the prediction. Likewise, the key to crystal structure prediction success is the ability to uncover the intricate details and hidden patterns that exist within crystal structures.

The previous chapter provided a clearer picture of advantages and drawbacks of existing crystal structure representation approaches. Upon this knowledge, we present in this chapter the features engineering process that we adopted for the transformation of raw crystal structure data into informative, numeric descriptors.

3.2 Experimental and computational data

Crystal structure data comes from two possible sources, namely experimental or computational. The former is data that we obtain from real measurements such as electrical, energetic, XRD, PND (powder neutron diffraction) ... etc. There are many different types of measurements depending on what we want to measure as properties on real materials that are synthesized in a laboratory, or that are found in their natural state (in situ).

Computational data, on the other hand, comes from an "in silico" experience (by means of computer modelling / simulation) on a computer, using codes that have managed to simulate physical and chemical reality by modeling the theory. The characterization of this data and its properties is performed using quantum mechanical techniques like DFT.

Take XRD data as an example; this data is experimental, but we can very well have a computationally simulated XRD. It is then to be compared with experimental data to validate the simulation. Indeed, in general, it is mandatory to validate our computational data with experimental data, provided the latter exists. It is absolutely possible to have purely computational XRDs, such as XRDs of pressurized terrestrial magma (inside the earth) that one cannot have in reality. However, to trust computational data that cannot be compared to experimental data, it is necessary to be rigorous in the theories of physics that are modeled in the computerized code. Both of these types of data are important in materials science as they complement each other to provide a comprehensive understanding of materials.

3.3 Features engineering

The extraction of features which may be used for model development is a crucial stage in the automated recognition of patterns and relationships from huge data sources. A feature, in general, represents a quality that was obtained from data input in its raw form in order to provide an appropriate representation. Therefore, features extraction seeks to identify elements of variation pertinent to the overall learning job and maintain discriminating information [96].

ML models strongly rely on these well-defined characteristics; the effectiveness of the extraction procedure will determine how well models perform. With time, a variety of features extraction methods that work with various data sources have evolved [96]. The features representing input data need to be suitable for ML modeling and must meet certain criteria.

The proposed features engineering process of the present work is illustrated in Figure 3.1.

3.3.1 Data collection

The data used in this study is of a computational type generated using USPEX [97, 98, 99] code. "Universal Structure Predictor: Evolutionary Xtallography" referred to as USPEX (pronounced as "uspekh", literally meaning "success" in Russian) is a method which was first developed in 2004 by the Oganov laboratory. USPEX is a computer program that is

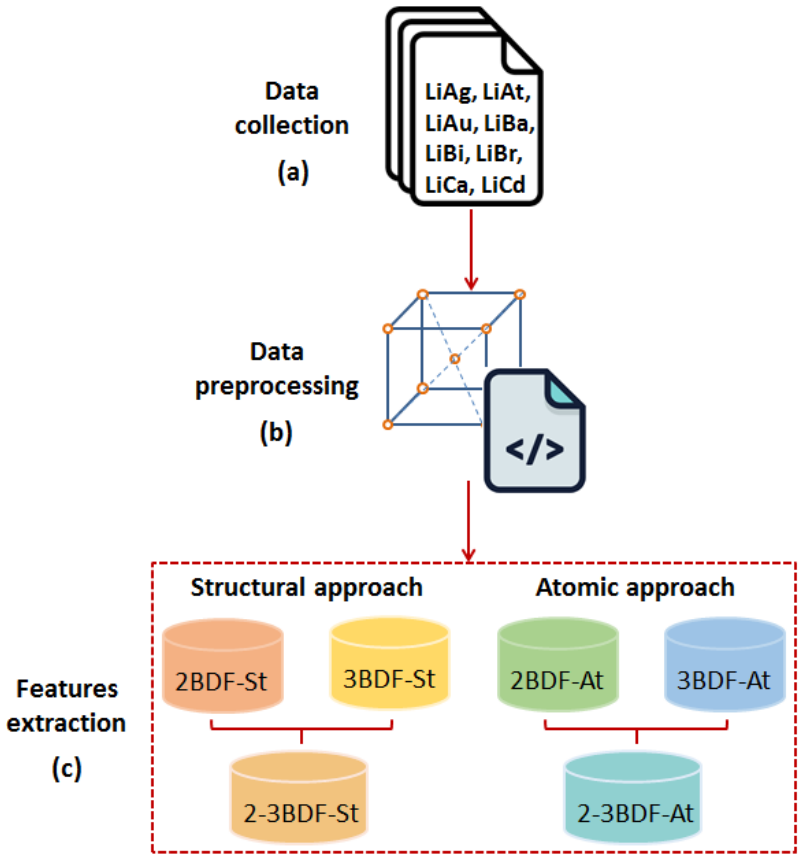


Figure 3.1: Proposed crystal structure features engineering process. (a) Collected databases, (b) data preprocessing, (c) extracted features categorized into six datasets.

Database	LiAg	LiAt	LiAu	LiBa	LiBi	LiBr	LiCa	LiCd	Total
Number of entries	1397	1379	1452	1404	1197	432	1247	1227	9717

Table 3.1: Investigated databases with their respective number of entries.

able to solve the crystal structure prediction problem under arbitrary pressure-temperature conditions using simply the material's chemical composition. It is based on evolutionary algorithm to search through a large number of possible crystal structures and find the most stable ones. In addition, it incorporates other external computational chemistry tools such as VASP, QUANTUM ESPRESSO, and Gulp.

USPEX was used to generate a total of eight databases that were considered for this study. As illustrated in Figure 3.1 (a), each of the databases comprises materials of one of these systems: Lithium Silver (LiAg), Lithium Astatine (LiAt), Lithium Gold (LiAu), Lithium Barium (LiBa), Lithium Bismuth (LiBi), Lithium Bromine (LiBr), Lithium Calcium (LiCa), and Lithium Cadium (LiCd). The number of instances in each of the aforementioned databases is given in Table 4.1.

3.3.2 Data preprocessing

The data extracted from USPEX (or other data sources) comes in its raw form. It is provided through a CIF (Crystallographic Information File) or a POSCAR file. Figure 3.2 represents a POSCAR file of the material Li_6Bi_2.

The POSCAR file presented in Figure 3.2 contains atom types of the crystal structure with their respective number as well as its geometric information. In addition, it has the output value of the target property to be predicted which, in this example, is the energy property with a value of -57.576002.

In order to perform machine learning modeling, one needs to select inputs and outputs and feed them separately to the model for the learning process. Therefore, it is required to identify the inputs from the outputs in POSCAR files. To this end, a preprocessing step was performed using simple NLP (Natural Language Processing) techniques.

3.3.3 Descriptors

In the case of crystal structures, features representing inputs are called descriptors [100, 101]. In order to have effective descriptors to be considered as machine learning model inputs, the features extraction process must insure that the data representation satisfies the following

```
EA17   7.790  7.790  7.790 90.00 90.00 9
OUT -57.576002
   1.00000000000000
     8.9890754595777977    0.0000000000000000    0.0000000000000000
     0.0000000000000000    6.1521741790493572    1.7218413129121517
     0.0000000000000000    1.5141608663974000    8.6559140081423340
   Li   Bi
    14     6
Direct
  0.7085196374237341  0.0001200000000026  0.0001200000000026
  0.2085196374237341  0.5001200000000026  0.0001200000000026
  0.3366206231128075  0.0215095634545435  0.3230282632499879
  0.8366206231128075  0.5215095634545364  0.3230282632499879
  0.8366206231128075  0.4787304365454546  0.6772117367500172
  0.3366206231128075  0.9787304365454617  0.6772117367500172
  0.3954883634126034  0.4314352161292661  0.6822523270200378
  0.8954883634126034  0.9314352161292661  0.6822523270200378
  0.1047454987041405  0.2217227890103928  0.8200216409659404
  0.6047454987041405  0.7217227890103928  0.8200216409659404
  0.8954883634126034  0.0688047838707320  0.3179876729799744
  0.3954883634126034  0.5688047838707320  0.3179876729799744
  0.6047454987041405  0.2785172109896052  0.1802183590340647
  0.1047454987041405  0.7785172109896052  0.1802183590340647
  0.3706959071562537  0.0001200000000026  0.0001200000000026
  0.8706959071562537  0.5001200000000026  0.0001200000000026
  0.1241377424804639  0.2881370480215537  0.4537887923378960
  0.6241377424804639  0.7881370480215537  0.4537887923378960
  0.6241377424804639  0.2121029519784443  0.5464512076621091
  0.1241377424804639  0.7121029519784443  0.5464512076621091
```

Figure 3.2: Example of a raw POSCAR file representing a data entry of the material Li_6Bi_2.

specific requirements [55]:

- Machine-readable representation.
- Complete: the representation must best describe a data without loss of information that is pertinent to the underlying issue.
- Compact: the representation should have the least redundant features.
- Unique and Nondegenerate: each data instance must have a unique representation and each representation must represent a single instance. For a representation to be unique and Nondegenerate, it must be invariant with respect to rotation, reflection, permutation and translation.
- Descriptive: two "close" instances having similar outputs must be represented through input features that are close in terms of distance.
- All representations of data in a dataset must be standardized and uniform.

Obviously, the example of crystal structure information given in Figure 3.2 cannot be used as a descriptor. Considering the fact that if a symmetry operation, such as rotation, is applied to the crystal structure, it would result in different atom coordinates for the same input, the information provided by the POSCAR file is not unique and nondegenerate. Moreover, extracting information from a POSCAR file, as it is, is not considered complete as there are many other data one could obtain using simple libraries such as Pymatgen [102]. In addition, materials differ in terms of atoms number which would make inputs not uniform.

In the previous chapter (chapter 2: State of the Art on Crystal Structure Prediction), an analysis on different crystal structure descriptors was investigated. Two important crystal structure descriptor-related conclusions were drawn, specifically: 1) PRDF is one of the most suitable descriptors for regression tasks (which is the type of prediction we are seeking in this work) and 2) MLIP-based approaches are dominant with regards to energy prediction (target property of our study) which implies that the descriptors used with these approaches are atom-wise ones.

Based on the aforementioned conclusions, the choice of crystal structure descriptors selected in this work is atom-wise distribution function-based descriptors as defined in [67]. The advantage with this choice is that it had been proven in previous studies that distribution function-based descriptors are effective. Moreover, atom-wise descriptors allow one to use MLIP modeling which has been demonstrated to be very accurate.

However, atom-wise descriptors are difficult to manage since they require developing a non-conventional machine learning topology. Therefore, we propose to investigate two types of distribution function-based descriptors, namely structural (structure-wise) descriptors and atomic (atom-wise) descriptors.

3.3.4 Two- and three-body distribution functions

As previously stated, machine learning interatomic potentials consider contributions of atoms in terms of energy as defined in the following equation.

$$E = \sum_{i=1}^{n} E_i \tag{3.1}$$

Such as E denotes the total energy of a material, E_i is the energy contribution of the i^{th} atom, and n is the number of atoms in the material.

In quantum interactions, two- and three-body interactions often account for the majority of the energy variation. If we focus on these two types of interactions, the energy would be formulated as follows:

$$E = \sum_{i<j} E_2(\vec{r}_i, \vec{r}_j) + \sum_{i<<kj} E_3(\vec{r}_i, \vec{r}_j, \vec{r}_k) \tag{3.2}$$

Where E_2 and E_3 are pair and triple interactions energies, i, j, k run through the material's atoms, and $\vec{r}$ are the atoms positions.

By considering the interatomic potentials that describe the interaction between two or more atoms, we take into account various physical and chemical factors that affect the bond between atoms, including the attraction between the nuclei and electrons, the distribution of electrons in the orbitals, and the repulsive forces between electrons. The bond between atoms is determined by the balance of these interatomic potentials. Stronger bonds result from a more negative interatomic potential, indicating a greater attraction between the atoms, while weaker bonds result from a less negative interatomic potential.

Similarly, the type of bond between two atoms affects the distance between them due to the forces of attraction between the atoms' nuclei and electrons. Stronger bonds, such as covalent bonds, result in shorter distances between the atoms, while weaker bonds, such as hydrogen bonds, result in longer distances. By taking distances (determined by the positions $\vec{r}$), we get the equation below:

$$E = \sum_{i<j} Q_2(|\vec{r}_i - \vec{r}_j|) + \sum_{i<j<k} Q_3(|\vec{r}_i - \vec{r}_j|, |\vec{r}_i - \vec{r}_k|, |\vec{r}_j - \vec{r}_k|) \tag{3.3}$$

Such as Q_2 and Q_3 represent two- and three-body potentials through one and three dimensional functions, respectively.

However, it should be noted that the complexity resulting from the summations of Equation 3.3 is $O(N^3)$, with N being the number of atoms. In order to reduce this complexity, two cut-off radii are introduced. The first one is R^2_{cut} denoting the limit distance that separates two atoms and above which all other pairs of atoms are ignored. The selected pairs of atoms meeting this criterion are denoted $P(R^2_{cut})$. Similarly, the triplets of atoms that are considered are defined using the second cut off radii R^3_{cut} through two variants: 1) a variant where all sides of the triangle forming a triplet of atoms do not exceed R^3_{cut}, and 2) a variant where the maximum length between at least two sides of the triangle is R^3_{cut}. Both variants' selected triplets of atoms meeting these criteria are designated as $T(R^3_{cut})$. This simplification results in an important decrease of the complexity to $O(N)$.

The right choice for R^2_{cut} and R^3_{cut} values depends on how much one would compromise the effectiveness of the descriptors with regard to the speed. Choosing a higher value for the cut offs would result in a higher accuracy but a slower potential and vice-versa.

The strategy chosen for this study in terms of descriptors is to represent data using distribution function-based potentials through a structural approach and an atomic approach. In these two approaches, descriptors based on two-body distribution function (2BDF), three-body distribution function (3BDF), and a combination of both two- and three-body distribution functions (2-3BDF) will be used. This strategy results in a total of six different descriptors, all invariant with respect to rotation, reflection, permutation, and movement, to be investigated, analyzed and compared for crystal structure data representation.

The features extraction process will be performed on the eight merged databases producing six databases each of nearly 10,000 entries. The six databases are obtained using 2BDF, 3BDF, and 2-3BDF of both (2D) structural and (3D) atomic approaches resulting in different features of the materials data.

3.3.4.1 Structural descriptors approach

In this approach, the descriptors are structure-wise, meaning that each structure is represented by a single descriptor as an input to which corresponds an energy value as a target. The mathematical representation of a structural descriptor is as follows:

$S_i : [v_1, v_2, \cdots, v_n]$

where S_i is the i^{th} structure in the dataset and v_j $(1 \leq j \leq n)$ is the j^{th} element of the descriptor vector of size n.

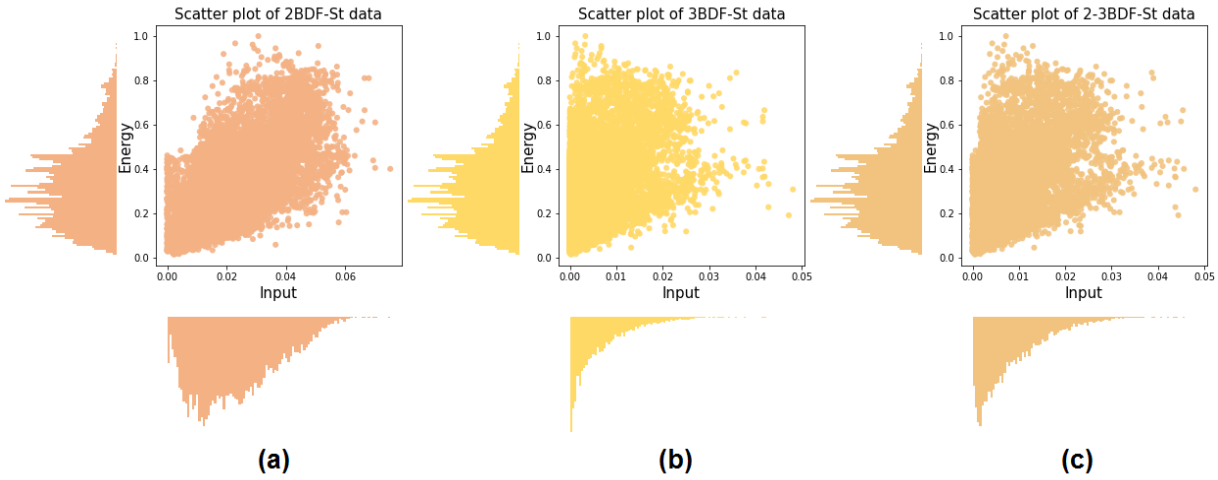


Figure 3.3: Overview of the three structural descriptors data distribution with the scatter chart of (a) 2BDF-St, (b) 3BDF-St, and (c) the 2-3BDF-St.

The size of the structural descriptor n depends on the applied data representation. Indeed, the number of elements of structural descriptor is either 60, 364, or 424 (60 + 364) for the structural two-body distribution function (2BDF-St), the structural three-body distribution function (3BDF-St), or the structural two- and three-body distribution functions combined (2-3BDF-St), respectively.

In order to make informed data-based decisions, it is essential to master the data at hand. One way to do that is through data visualization. By the use of different graphs and plots, large and complex data can be made easier to understand while allowing more meaningful insights. Therefore, we proceed hereafter to data visualization of the three structural descriptors 2BDF-St, 3BDF-St, and 2-3BDF-St.

Figure 3.3 represents the scatter plot of the eight databases features combined and represented through (a) the 2BDF-St descriptor, (b) the 3BDF-St descriptor, and (c) the 2-3BDF-St descriptor.

A scatter plot of input features against output in data visualization can be very practical in terms of revealing the relationship between the input features and the output. It can expose the power and direction of the relationship, as well as eventual non-linearities. The scatter plots in Figure 3.3 (a), (b), and (c) illustrate data points of the 2BDF-St, 3BDF-St, and 2-3BDF-St descriptors, respectively, as the input data average against the corresponding normalized energy output value.

As opposed to a linear relationship between inputs and outputs, which is illustrated by

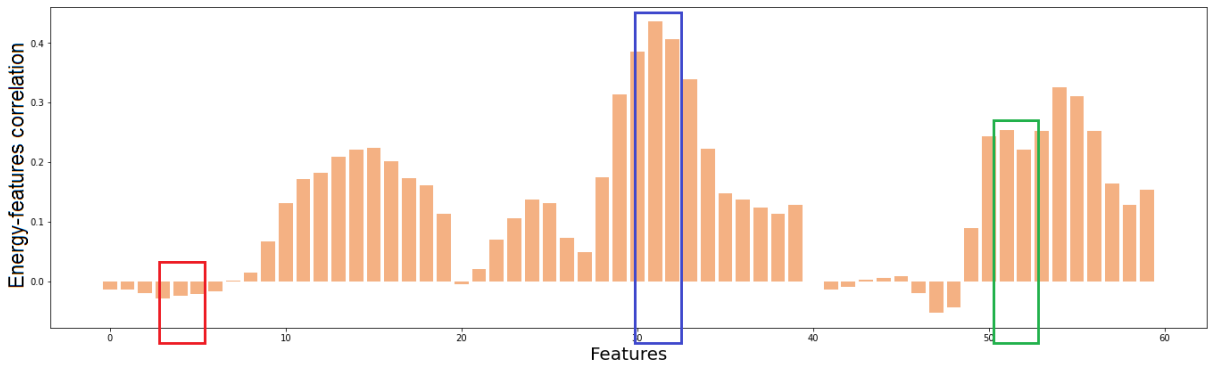


Figure 3.4: Bar chart illustration of the 2BDF-St features' correlation with the energy property.

a clear pattern of points forming a straight line in the scatter plot, a non-linear relationship is represented by a curved pattern or clusters of points. From the plots in Figure 3.3, a non-linear behavior is clearly discernible by clusters of points. The data point distribution is concentrated in the intervals [0.005-0.035, 0.1-0.5], [0-0.005, 0.1-0.5], and [0-0.01, 0.1-0.5] for 2BDF-St, 3BDF-St, and 2-3BDF-St descriptors, respectively, while it's lightly scattered outside these intervals. Compared to 2BDF-St descriptor scatter, the 3BDF-St one is even less proportionate, and thus, less linear, while the third structural descriptor is a combination of the former two. It is constructed by the horizontal concatenation of 2BDF-St and 3BDF-St and therefore has the characteristics of both.

In order to make sure that the data representation of 2BDF-St, 3BDF-St, and 2-3BDF-St descriptors is adequate for crystal structure energy prediction, the correlation between the inputs and the corresponding outputs needs to be analyzed. In machine learning, this correlation refers to the relationship between the input features of a data set and the target variables to be predicted. The strength of the correlation between inputs and outputs affects the ability of an ML model to generalize and make effective accurate predictions; the stronger the correlation, the better the model performance. Figures 3.4, 3.5, and 3.6 illustrate the correlation between 2BDF-St, 3BDF-St, and 2-3BDF-St descriptors features of the investigated dataset, respectively, with the energy value.

As depicted in Figures 3.4, 3.5, and 3.6, the features of 2BDF-St, 3BDF-St, and 2-3BDF-St descriptors of sizes 60, 364, and 424 have different correlation values with the energy, ranging from [-0.053, 0.4358], [-0.0307, 0.4009], and [-0.053, 0.4358], respectively. The 3BDF-St correlation interval being smaller than that of 2BDF-St descriptor confirms the stronger non-linearity. The 2-3BDF-St features-energy correlation interval is the same as

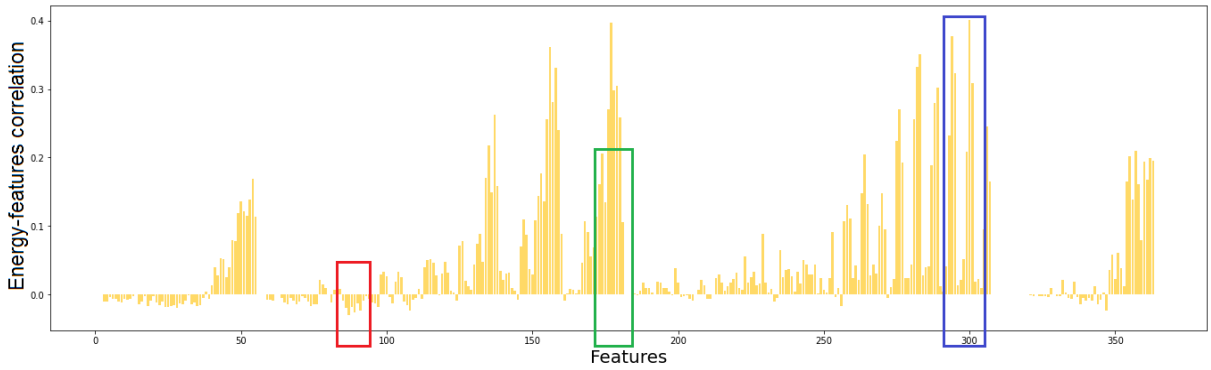


Figure 3.5: Bar chart representing the 3BDF-St features' correlation with the output energy property.

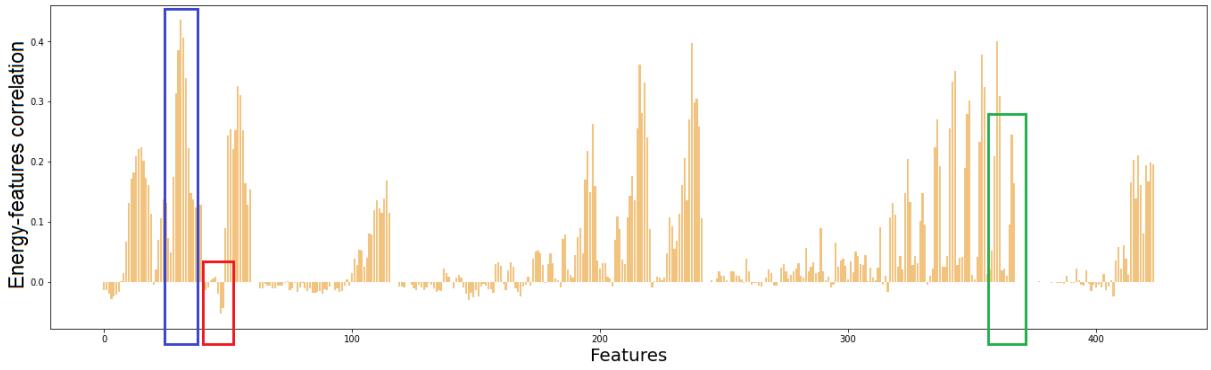


Figure 3.6: Bar chart depiction of the correlation between the 2-3BDF-St features and the energy property.

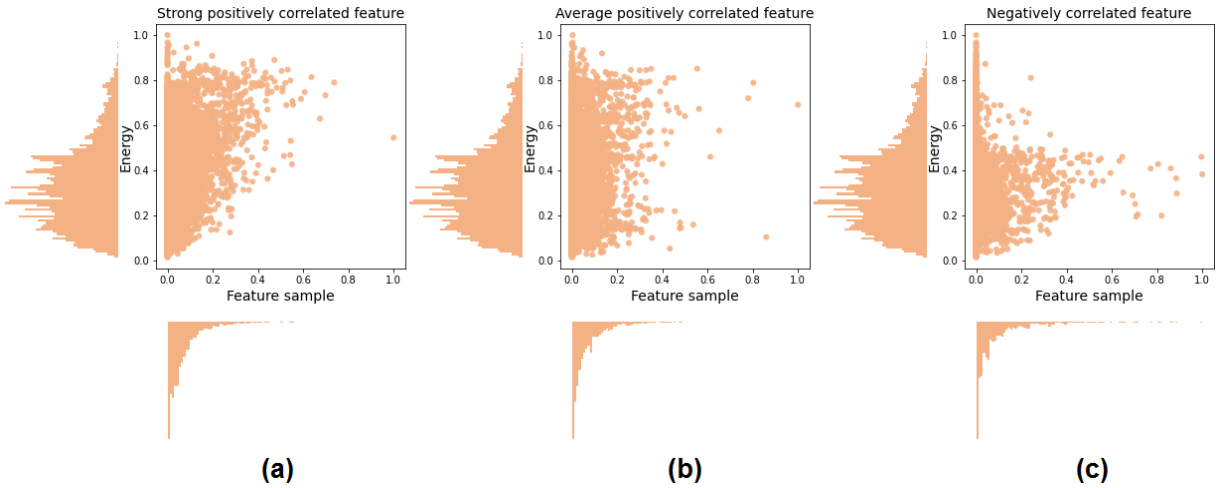


Figure 3.7: Scatter plot of three features samples of the 2BDF-St descriptor. (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

that of 2BDF-St descriptor since it is larger and includes the 3BDF-St one. It is to be noted that negative correlations are as important as positive ones. Where a positive correlation refers to the fact that, as the values of the input increase, those of the output variable also increase, a negative correlation refers to the opposite relationship, i.e. an increase in the input causes a decrease in the output. Since it is difficult and not appropriate for this data to be fully graphically presented, the choice of a normalized-based sampling representation of the data was adopted by selecting specific samples of features based on their correlation with the energy. To better visualize the non-linearity of data, we chose to plot three features samples, namely: 1) a strong positively correlated feature highlighted by a blue frame, 2) an average positively correlated feature highlighted by a green frame, and 3) a negatively correlated feature highlighted by a red frame in Figures 3.4, 3.5, and 3.6. These features samples' scatter plots are presented in Figures 3.7, 3.8, and 3.9.

We can clearly notice that each sample scatter of one of the three descriptors differs from the others in terms of data distribution. In addition, the three sub-plots from Figures 3.7, 3.8, and 3.9 further prove the data non-linearity behavior. Moreover, although the data distribution is non-linear and disproportionate, we notice that all the randomly selected samples representing positive correlation features form the same pattern shape for the three descriptors which appears like a "V" shape. Likewise, the chosen samples that indicate a negative correlation form a similar pattern appearing like an "A" shape (or an inversed "V")

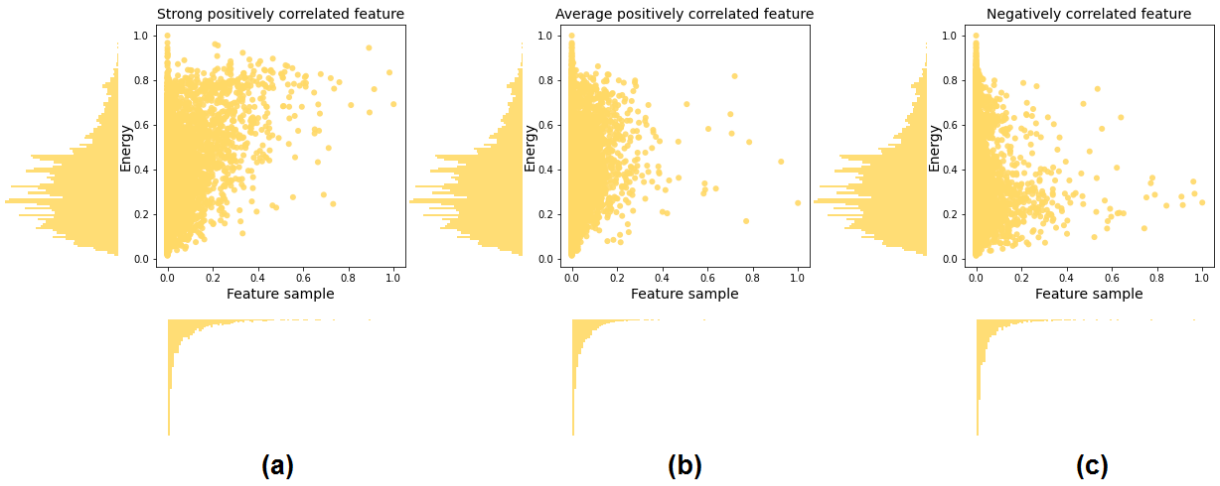


Figure 3.8: Scatter plot of three features samples of the 3BDF-St descriptor. (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

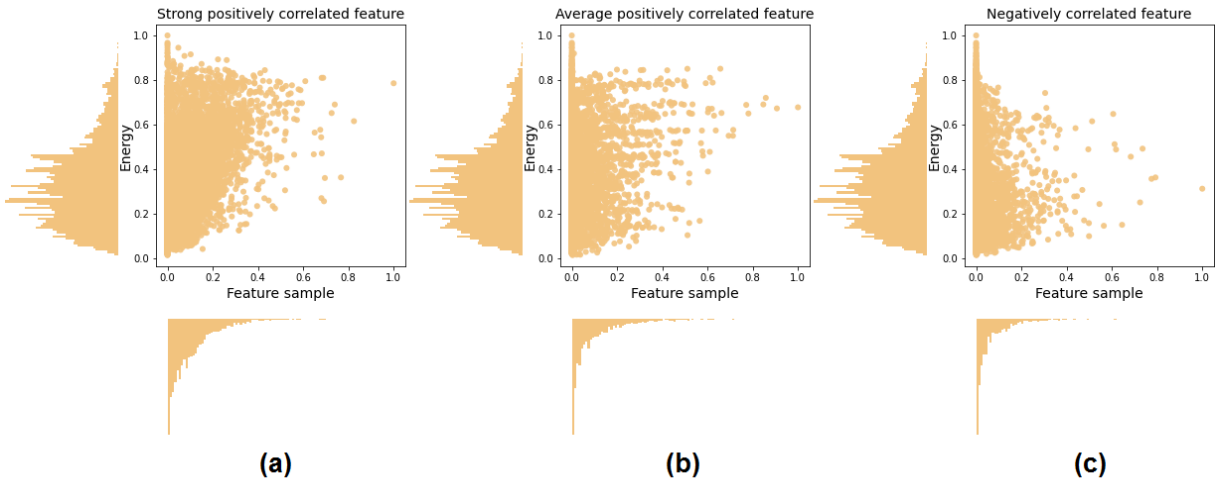


Figure 3.9: Scatter plot of three features samples of the 2-3BDF-St descriptor. (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

mirroring the opposite relationship of a positive correlation between the input features and the output data.

3.3.4.2 Atomic descriptors approach

Unlike the structural approach, in the atomic approach, each atom has its own descriptor vector, resulting in a structure with as many descriptor vectors as the number of atoms. An atom-wise descriptor for a given structure S_i is mathematically represented as follows:

$$S_i : \begin{bmatrix} v_{11}, & v_{12}, & \cdots, & v_{1n} \\ v_{21}, & v_{22}, & \cdots, & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1}, & v_{m2}, & \cdots, & v_{mn} \end{bmatrix}$$

Where S_i is the i^{th} structure in the dataset and $v_{jk}(1 \leq j \leq m, 1 \leq k \leq n)$ is the k^{th} element of the descriptor vector of the j^{th} atom, with n and m representing the size of the descriptor vector and the number of atoms in structure S_i, respectively.

Similar to the structural approach, three different atomic data representations are used. The number n representing the size of an atomic descriptor depends on which data representation is applied. According to whether an atomic two-body distribution function (2BDF-At), an atomic three-body distribution function (3BDF-At), or the combination of the two former ones as an atomic two- and three-body distribution function (2-3BDF-At) is utilized, the number of elements of the atomic descriptor is either 60, 468, or 528 (60 + 468), respectively.

For this atomic approach, we use the same data visualization strategy to reveal and illustrate the relationship between 2BDF-At, 3BDF-At, and 2-3BDF-At descriptors with the energy output. Indeed, investigating the relationship between the input features and the output is an important step in the machine learning process, as it can help to boost the performance and interpretability of the model, and more importantly, identify any potential issues with the data or model.

In Figure 3.10 (a), (b), and (c), we proceed to depict the scatter plot of the input features represented through 2BDF-At, 3BDF-At, and 2-3BDF-At descriptors, respectively.

The scatter plots in Figure 3.10 show the data distribution of the three different atomic descriptors. We notice that, from a plot to another, the scatter pattern changes. The data distribution is concentrated in the intervals [0.012-0.075, 0.1-0.5], [0-0.01, 0.1-0.5], and [0-0.03, 0.1-0.5] for the descriptors 2BDF-At, 3BDF-At, and 2-3BDF-At, respectively. As the scatter appears like clusters of points, the relationship between the input features represented through the three atomic descriptors with the energy output is non-linear.

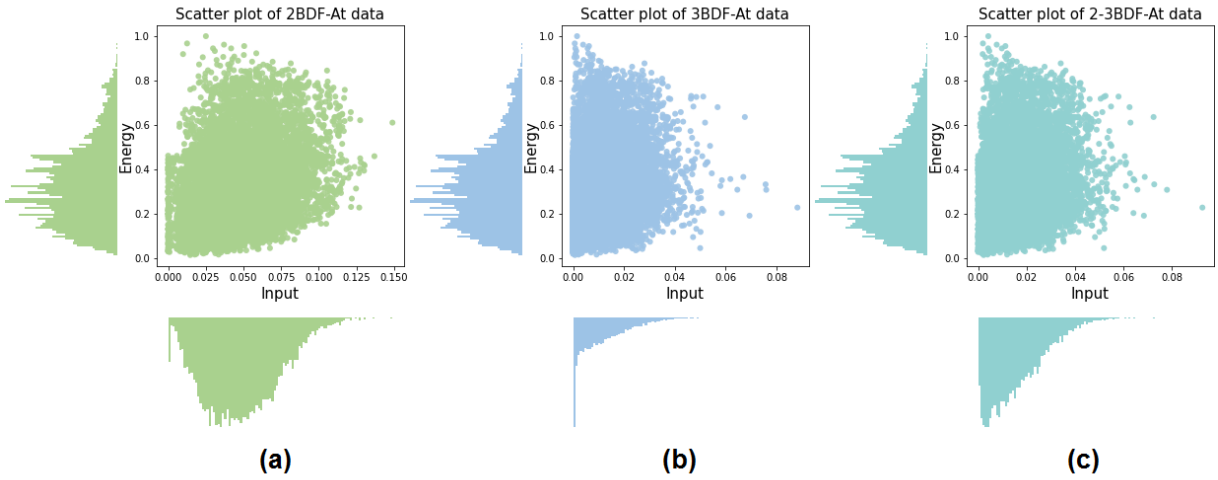


Figure 3.10: Overview of the three atomic descriptors data distribution with the scatter chart of (a) 2BDF-At, (b) 3BDF-At, (c) 2-3BDF-At.

To further examine this relationship, we investigate the correlation of the data features extracted using 2BDF-At, 3BDF-At, and 2-3BDF-At descriptors with the energy in Figures 3.11, 3.12, and 3.13, respectively.

The Figures 3.11, 3.12, and 3.13 depict the output-features correlation using 2BDF-At, 3BDF-At, and 2-3BDF-At descriptors of size 60, 468, and 528, respectively. The correlations of the three descriptors 2BDF-At, 3BDF-At, and 2-3BDF-At with the energy range between [-0.053, 0.4358], [-0.0241, 0.4425], and [-0.053, 0.4425], respectively. In each of these three correlation plots, we highlighted three features samples in blue, green, and red representing a strong positive, an average positive, and a negative correlation with the energy output, respectively. These correlations are to be inspected in Figures 3.14, 3.15, and 3.16.

We observe in the strong positively correlated feature plots of 2BDF-At, 3BDF-At, and 2-3BDF-At the same behavior as for the structural descriptors; i.e. the data distribution forms a pattern in the shape of a "V". Likewise, negatively correlated feature plots show the reversed behavior illustrated by a data distribution pattern arranged in the shape of an "A" (or an inversed "V"). This proves that the various types of correlation between the input data of the three different descriptors are consistent and show no discrepancy.

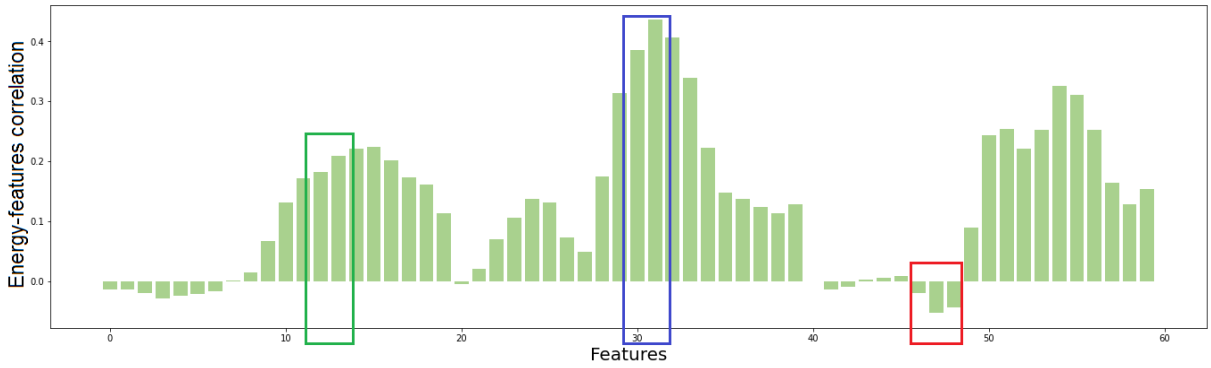


Figure 3.11: Bar chart representing the 2BDF-At features' correlation with the output energy property.

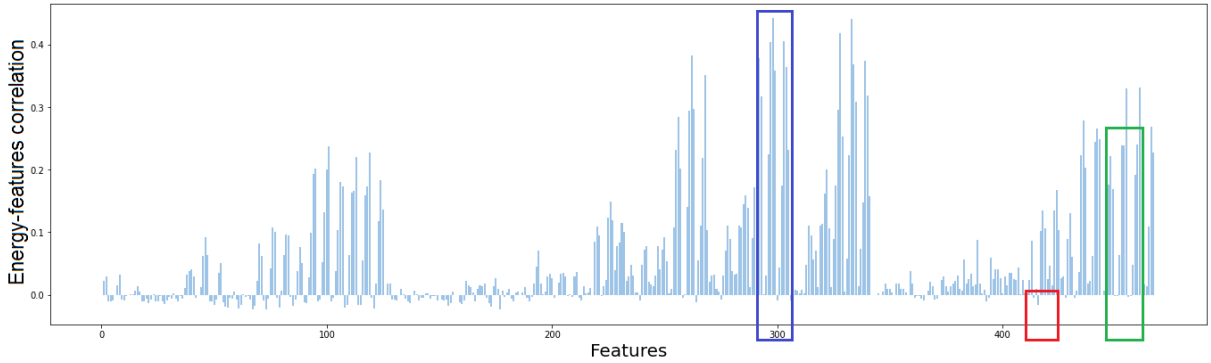


Figure 3.12: Bar plot illustration of the correlation between the 3BDF-At features and the energy property.

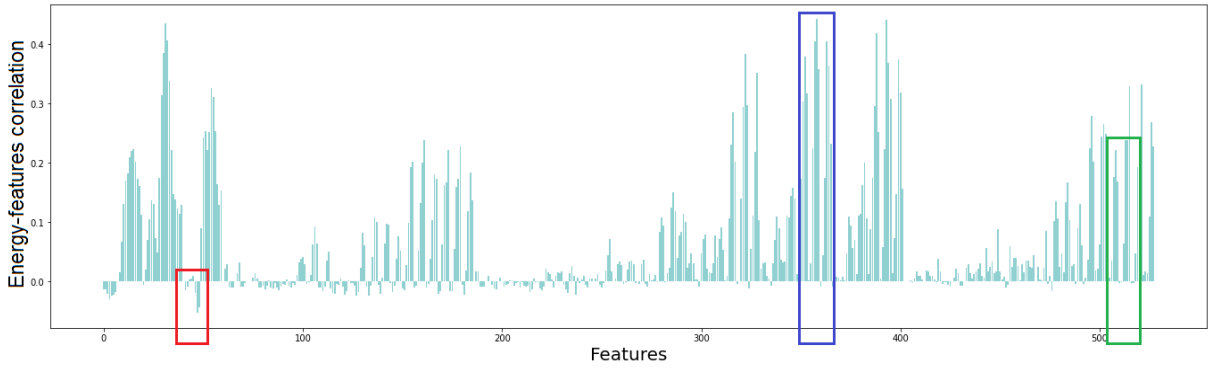


Figure 3.13: Bar plot depiction of the 2-3BDF-At features' correlation with the output energy property.

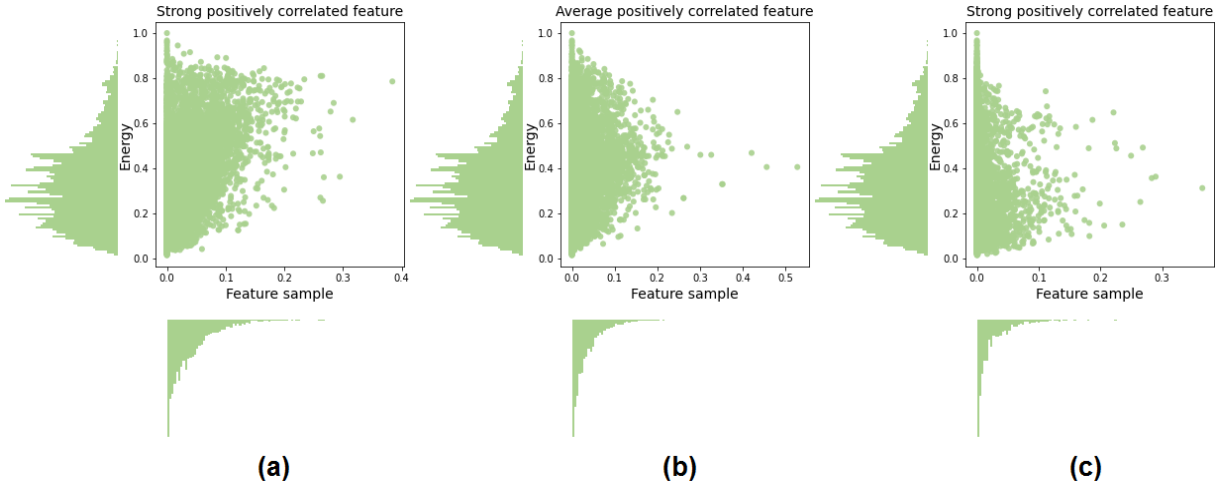


Figure 3.14: Scatter plot of three features samples of the 2BDF-At descriptor, (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

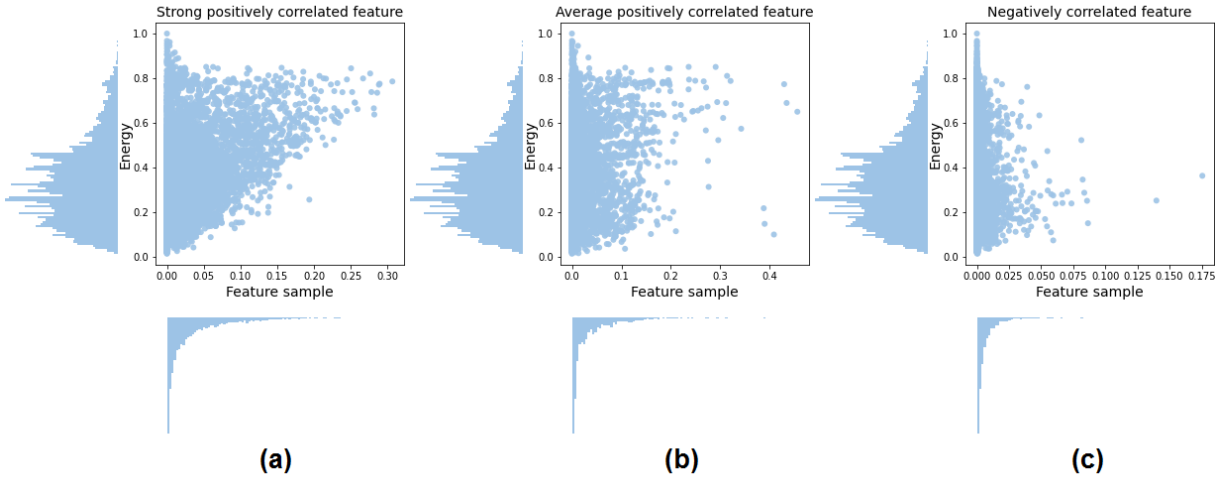


Figure 3.15: Scatter plot of three features samples of the 3BDF-At descriptor, (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

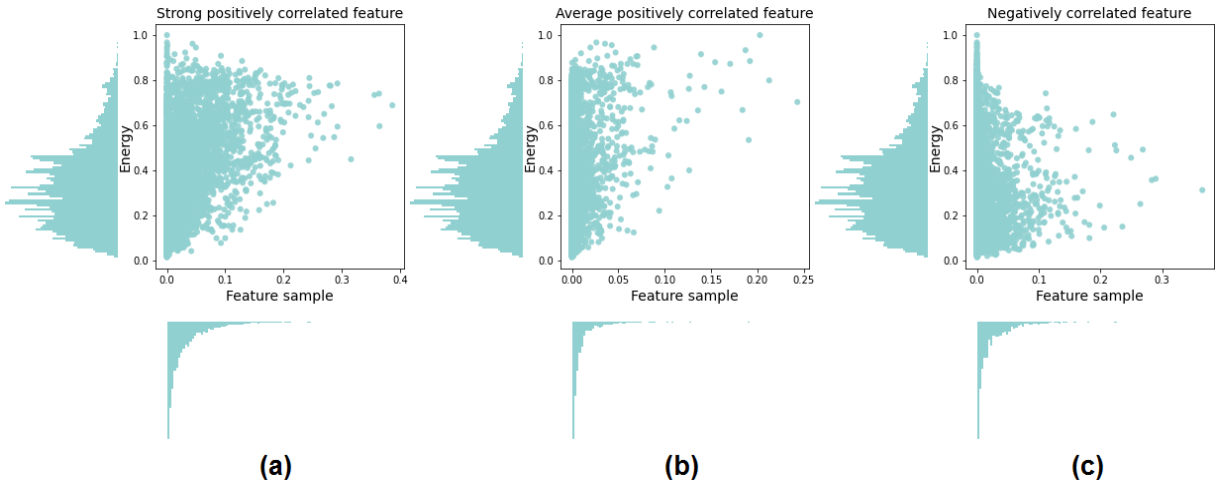


Figure 3.16: Scatter plot of three features samples of the 2-3BDF-At descriptor. (a) Strong positively correlated feature, (b) average positively correlated one, (c) negatively correlated one.

3.4 Summary and discussion

The significance of examining the correlation between the input features and the output can help to acquire understanding of the fundamental relationship between the input and output; thus, making it simpler to interpret the model's predictions. Moreover, analyzing the correlation between the input features and the output can help spot overfitting and prevent it. Indeed, if the input features are highly correlated with the output, the ML model may overfit the testing data and underperform on new, unseen data. In contrast, if the input features are weakly correlated with the output, it may indicate that more features or additional rigorous features engineering methods are required to boost the performance of the ML model. In our case, the highest correlation recorded is 0.4425 which is very a suitable correlation value since it is more likely to be moderate than high or weak.

In summary, the response of all plots illustrates the non-linear relationship nature where the output (energy) is indirectly proportional to the input (features), and any change in the input produces a disproportionate change in the output, as evidenced by energy plotted versus features properties using both structural and atomic representations.

3.5 Conclusion

This chapter has unveiled the need of engineering significant features and shown how the raw crystal structure data can be transformed into an ML-suitable, numeric representation. The two- and three- body distribution functions have been used for this purpose with structural and atomic approaches.

As we venture forward, the conclusions drawn about the nature of the relationship between the dependent and independent data, as well as the understanding gained from the CSP state-of-the-art, are considered a valuable foundation for informed decision-making in the modeling stage which will be covered in the next chapter.

Chapter 4

Proposed Crystal Structure Energy Prediction Modeling with Machine / Deep Learning

"Predicting the future is not magic, it's artificial intelligence." – Dave Waters.

4.1 Introduction

With its ability to identify complex patterns within data, machine learning has become an essential tool in the search for predictive accuracy. In the field of materials science, the fusion of data and learning has the potential to solve the riddles of crystal structures. The present modeling stage focuses on creating prediction models that give the insights acquired from the features engineering process a deeper significance. The meticulously engineered features from the previous chapter now serve as inputs for our prediction models. As the connections between crystal structures and the desired energy property are uncovered, prediction models that harness the power of data are built.

In this chapter, following the features engineering course, we will carry on with the crystal structure energy prediction through two approaches: structural modeling with machine/deep learning using the structural descriptors and atomic modeling with deep learning using atomic descriptors.

4.2 Machine learning predictions

The machine learning method employs algorithms made to identify and learn patterns in data and provide predictions using that learning. The general mechanism of machine learning consists in 1) reading input data that is represented by a number of characteristics (features), 2) training a machine learning algorithm to discover patterns or a data structure, or to learn the correlations between the characteristics in the input data and the target variable, 3) adjusting the parameters of the model that determine the weight and intensity of certain features in a way that minimizes the learning error, and 4) proceeding to the prediction process on new, unseen data using the built model that has been trained on the input data.

Depending on the prediction type, machine learning modeling is divided into supervised learning and unsupervised learning [103]. Unsupervised learning models are only provided with unlabeled data; their objective is to detect patterns or structure in the data, such as grouping instances together based on similarity or determining underlying variables that explain the data [104]. Clustering is one among various unsupervised learning techniques with the aim of arranging and grouping the data into clusters so that the instances inside each cluster are similar to one another and distinct from other clusters' instances [105].

Supervised learning, however, is a type of machine learning where the model is trained on data which is labeled with the target value, and the objective is to predict new, unforeseen instances using the patterns discovered from the training data. In supervised learning, every instance in the training data has a label or target variable associated that denotes the desired result. The model is trained on a collection of (x, y) pairs, where x represents the independent variable and y the dependent one, to construct a sort of a complex mathematical function which is able to predict the target variable $y*$ in response to a query $x*$ as a new instance [106].

Supervised learning can be further divided into two sub-categories: classification and regression. The former consists in predicting a categorical target variable given a set of features [107]. For example, in the medical field, predicting whether a patient has diabetes or not based on a set of information such as age, gender, weight, etc. is a binary classification problem, i.e. a prediction problem with two possible outcome classes. There's also a multi-classification problem where the number of outcome classes exceeds 2. A famous practical case for multi-classification task is the prediction of digits based on handwritten digits image pixel values.

The goal of regression, on the other hand, is to predict a continuous quantitative target variable given a set of features. For example, given data on housing prices, the goal might

be to predict the price of a house based on its size, location, year of construction, and other factors [108].

In general, regression algorithms are used in a wide range of applications including materials science, social media, and marketing, to name a few. There are numerous types of regression models; the choice of a regression model is determined by the nature of the problem at hand and the characteristics of the data. For instance, linear regression is suitable for data with a matching simple linear input-target variables relationship, while decision tree regression is more appropriate for data with non-linear relationships and complex interactions between the input features.

In this study, we seek to predict the energy property, a continuous quantitative value, of crystal structures that are represented through structural and atomic two- and three-body distribution functions. These descriptors, as seen in the previous chapter, have a non-linear relationship with the target variable.

Since the dataset at hand consists of eight databases in which each instance is labeled with the energy output value, the problem we face herein is a supervised non-linear machine-learning-based regression.

We proceed hereafter to the modeling stage of the energy property prediction through structural and atomic approaches using respectively structural and atomic two- and three-body distribution functions.

4.3 Structural approach modeling

In the modeling stage of the structural approach, we use the structural two- and three-body distribution functions (2BDF-St, 3BDF-St, and 2-3BDF-St) as inputs. As seen in the previous chapter, the structural data is suitable for machine learning where each structure in the database is represented through a vector of 60, 364, or 424 depending on the used descriptor. Therefore, we can proceed to the modeling by selecting appropriate regression models to predict the energy value, as depicted in Figure 4.1.

The input data extracted from 2BDF-St, 3BDF-St, and 2-3BDF-St is a complex data with different levels of non-linearity. When selecting machine learning algorithms for the modeling stage, one would trivially select models according to the nature of the input data; i.e. linear models for linear data and non-linear models for non-linear data. However, it is sometimes interesting to explore more and think outside the box. In case of linear data, the problem does not arise, since linear models are well suited and enough to solve the problem.

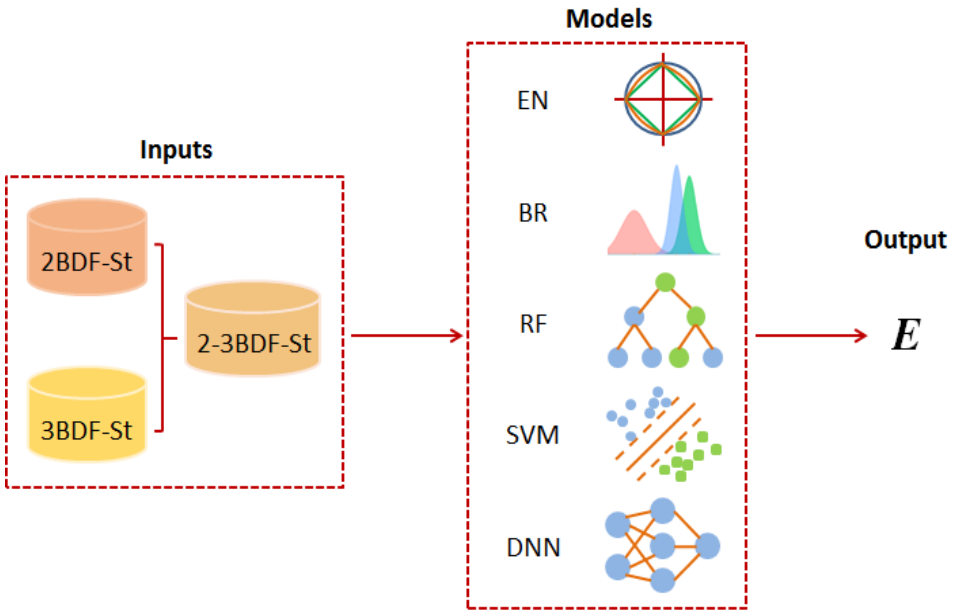


Figure 4.1: Structural approach modeling for the prediction of the energy property. (a) Inputs, (b) selected machine/deep learning algorithms, (c) output.

In fact, choosing a non-linear model for linear data would most probably cause overfitting. If not, it would simply be overkill, thus wasting energy and resources.

On the other hand, when data is non-linear, there is room for inspection. It is true that linear models assume that the relationship between the input and the output variables is linear. This means that the output variable can be expressed as a combination of the input variables with some coefficients that is of a linear nature. These coefficients essentially govern the strength (or weight) and direction of the relationships. However, linear models can sometimes work surprisingly well on non-linear data. This might be the case when non-linear relationships may not be very strong. Indeed, when the non-linear relationships between the input and output variables are relatively weak, a linear model can nevertheless account for the majority of the variance in the data. A linear model can also capture certain non-linearities in the data, in case the input variables are highly correlated with each other, or in case the features engineering process has transformed input data that was originally linear into new features that have non-linear relationships with the output variable.

The main reason one would first reach out to linear solutions is the computational cost. Indeed, linear models are simpler algorithms which do not require a lot of data preprocessing such as scaling or normalization. Such procedures might very well be computationally

expensive in case of a large dataset. Moreover, linear algorithms have fewer parameters to estimate than non-linear ones; thus, decreasing the complexity and the computational effort to fit the model. In addition, when a linear model is built and trained, making off-line predictions is typically computationally less costly than for non-linear models. This is because, in contrast to non-linear models, which frequently necessitate more sophisticated computations, non-linear models just require a matrix multiplication as the computation needed to generate a prediction.

In the following sub-sections, we will present the theory of the investigated machine learning models that were considered for the structural approach modeling. We will start with linear models and make our way to non-linear ones which proved their efficiency in the state of the art.

4.3.1 ElasticNet

ElasticNet (EN) is a linear machine learning algorithm which uses linear regression. The linear regression algorithm seeks to define the relationship between the input and output variables by finding the best-fit line through the data. The best-fit line of a simple linear regression (having only one independent variable) is defined as follows (see equation below):

$$y = \beta_0 + \beta x \tag{4.1}$$

Where y is the dependent variable (to be predicted), x is the independent variable (the features), β is the slope of the line (how steep the line is), and β_0 is the y-intercept of the line. The main goal of linear regression is to find the values of β_0 and β in such a way that the difference (error) between the predicted values of y and the actual y target values is minimized.

As opposed to simple linear regression, multiple linear regression is when the input feature is composed of two or more variables. In this case, the equation above becomes as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{4.2}$$

Where $x_i, 1 \leq i \leq n$ with n being the number of features, are the independent variables and β_i are the regression coefficients determining the weight of each independent variable (variable contribution to the value of y). In order to find the best-fit line, one needs to determine β_i in a way to minimize the cost function defined in the equation below [109]:

$$\hat{\beta} = argmin \sum_{i=1}^{m} (y_i - (\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}))^2 \tag{4.3}$$

However, linear regression as defined might suffer from several drawbacks, such as model complexity in case of a dataset with a large number of features and overfitting. In order to fix these problems, two linear regression-based algorithms are introduced, namely: LASSO (least absolute shrinkage and selection operator) and ridge regression. In LASSO regression, a penalty term (L1 regularization) is added to the linear regression's cost function (see Equation 4.4). This penalty pushes the model to shrink the least important features' coefficients to zero [109].

$$\hat{\beta}_L = argmin(\sum_{i=1}^{m} (y_i - (\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}))^2) + \lambda \sum_{j} |\beta_j| \tag{4.4}$$

Where the λ parameter controls the strength of the regularization; a larger lambda value implies that more coefficients are shrunk towards zero and thus irrelevant features removed.

Compared to linear regression, LASSO regression has the advantage to cope with high-dimensional datasets having many features. Also, by only focusing on the most crucial features for prediction, it can produce a model that is easier to understand.

Ridge regression is also a linear regression-based algorithm. Its most important advantage is avoiding overfitting. Similar to LASSO regression, a penalty term (L2 regularization) is added to the cost function of linear regression as shown in Equation 4.5 below [109]:

$$\hat{\beta}_R = argmin(\sum_{i=1}^{m} (y_i - (\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}))^2) + \lambda \sum_{j} (\beta_j)^2 \tag{4.5}$$

One way to benefit from both LASSO and ridge regressions advantages is to use ElasticNet ML algorithm. Indeed, EN is an algorithm based on linear regression that couples LASSO and ridge regressions. Given that it combines the ridge and LASSO regression techniques, the ElasticNet plot, when displayed on a Cartesian plane, lies between them.

Mathematically, by combining the penalties of both L1 and L2 regularizations, EN can produce more accurate predictions than either L1 or L2 regularization alone. Its estimator is defined in the following equation [109]:

$$\hat{\beta}_{EN} = argmin(\sum_{i=1}^{m} (y_i - (\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}))^2 + \lambda(\alpha \sum_{j} |\beta_j| + (1-\alpha) \sum_{j} (\beta_j)^2)) \tag{4.6}$$

Where α represents a tuning parameter that can regulate the ratio of L1 and L2 regularizations. This parameter, along with λ, allow for fine-tuning of the model.

Therefore, EN model can benefit from both LASSO and ridge regressions' strengths which are respectively known to be 1) removing low relevance features and 2) reducing the chance of data overfitting [110].

4.3.2 Bayesian Ridge (BR)

Bayesian ridge is a model combining both the Bayesian regression and the previously explained ridge regression. The Bayesian regression is a statistical approach based on Bayes' theorem. This method brings interesting assets to the linear regression as it can take into account interactions between inputs variables and, more importantly, nonlinear relationships between the independent variables and the response variable. Plus, one can rely on Bayesian regression to account for uncertainty.

Starting from the following fundamental equation of a linear regression model [111]:

$$y = \beta X + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{4.7}$$

where y and X are respectively the response and predictor variables, β the weight coefficients, and ϵ an error exhibiting a normal distribution with a mean of 0 and variance σ^2. What a Bayesian regression model tries to predict is the parameter vector β (weight coefficients which can solve Equation 4.7) based on some observed data y. This is called the posterior distribution and it's illustrated in Equation 4.8.

$$\hat{\beta} = argmax(P(\beta|y)) \tag{4.8}$$

By applying Bayes theorem, we obtain the following equation:

$$\hat{\beta} = argmax(\frac{P(y|\beta) \times P(\beta)}{P(y)}) = argmax(P(y|\beta) \times P(\beta)) \tag{4.9}$$

In Equation 4.9, the denominator $P(y)$ was ignored because it has no relation with β, which leaves us with the likelihood (1^{st} term of the equation) and the prior distribution (2^{nd} term of the equation). This Bayesian-based method primarily allows for any prior knowledge, information or beliefs that one may have on the model's parameters to be reflected [112].

A variety of Bayesian regression called Bayesian ridge regression utilizes a ridge penalty on the model coefficients in order to incite shrinkage towards zero. The ridge penalty may be viewed as a method of accounting for the prior belief that the coefficients should be small unless there is compelling evidence to the contrary.

In Bayesian ridge, the parameters' prior distribution is modeled with a Gaussian distribution with a mean of 0 and λ. The parameters' prior knowledge is represented by the precision λ which also regulates how much shrinkage is applied to the coefficients. In practice, the Bayesian ridge model allows to control both the L1 and L2 regularizations for a stronger performance.

4.3.3 Random forest

Decision trees are simple, straightforward statistical-based algorithms that can perform both classification and regression tasks accurately [113, 114]. The primary feature of decision trees is the recursive partitioning of datasets into descendant data subsets based on the values of associated predictors [115]. The general algorithm of a decision tree is composed of three main steps. The first step consists in selecting a feature that most effectively divides the data into two subsets. For this step of the algorithm, the key element is to best choose the feature for splitting the data. To this end, in case of classification, the feature is selected using metrics such as information gain that measures how much the uncertainty is reduced (the higher the value of information gain the better) [116] or Gini impurity which typically calculates the probability of a randomly chosen data being misclassified (the lower the value of Gini impurity the better) [117].

In case of regression however, this step consists in choosing a split point on a certain feature in a way to best split the data into subsets. The split point can be any value ranging from the minimum to the maximum value of the selected feature x_i. The metric used for this selection is of an error type such as Mean Squared Error or Mean Absolute Error (the lower the value of the error measure the better). Once the feature and split point have been selected, the second step of the algorithm is to divide the data of the current node into subsets according to the chosen feature and split point. Then, steps 1 and 2 are repeated until a stopping criterion is met. This criterion represents the third step of the algorithm. Each time new subsets are created through step two's data splitting, the stopping criterion is tested to whether it is met or not. It can be chosen to be defined by the minimum number of samples in a leaf node, maximum tree depth, etc.

Random forest [118] is a decision tree-based ensemble method [119]. It is applied for classification as well as regression problems using multiple decision trees for the prediction process. It consists of an ensemble of n decision trees where each tree $T_i(X), i = 1, \cdots, n$ with $X = x_1, \cdots, x_m$ being the features vector of a data, is generated using a random vector that is independent from the input vector [120].

The training algorithm of random forest is the following [121]:

1. Draw bootstrap samples from the training data.
2. For each sample, grow an unpruned tree where each node consists of the best chosen split among a randomly selected subset of descriptors.
3. Predict new data by aggregating (average for prediction) the predictions from all trees.

The prediction of random forest produces n outputs (one per each tree) and the final output is the aggregation of all trees' outputs [122]. The random forest prediction is defined in Equation 4.10 as the unweighted average over the tree collection [123].

$$h(x) = \frac{1}{N} \sum_{i=1}^{N} h(x, \Theta_i) \tag{4.10}$$

Where x stands for the observations with associated random vector, and Θ_k represents independent and identically distributed random vectors.

The generation of bootstraps sample, the features subsets that are randomly chosen, and the prediction aggregation are the three key components that make a random forest model robust. Although random forests are typically more complex and less interpretable compared to single decision trees, they provide more accurate predictions, less overfitting, and noise resilience [124].

4.3.4 Support vector machine

Support vector machine is an ML model that is mainly used for binary classification problems. Nevertheless, the general principle of SVM can be applied the same way on a regression task. Indeed, since the relationship between a multidimensional input vector x and the output is most likely to be non-linear, it is necessary for the data features to be mapped into a high dimensional space in order to create a linear hyperplane [125] by the use of an SV kernel [126]. Then, rather than using the hyperplane as a decision boundary in a classification task to distinguish between patterns, SVM (also referred to as SVR short for Support Vector Regressor when applied to regression problems) looks for a match between the input vector and its position in the curve in order to predict a continuous real value. Its main goal is to construct the hyperplane such that it lays close to the data points. This translates to choosing a hyperplane with small norm while the sum of the distances from the hyperplane to the data points is minimized [127].

Given the training points (x_i, y_j), the general form of the hyperplane is defined by the equation below [125]:

$$f(x) = wx + b \tag{4.11}$$

The regularized risk functional which needs to be minimized is defined as:

$$min\frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} L_\epsilon(y_i, f(x_i)) \tag{4.12}$$

The first term is the regularization term that is related to the model complexity as explained in Cortes and Vapnik [128]. In the second term, the loss function L_ϵ ignores the prediction error if the difference between the predicted value $f(x_i)$ and the actual value y_i is smaller than ϵ [125]; and $C \geq 0$ represents the tradeoff between prediction accuracy and cost.

We have $y_i - wx_i - b \leq \epsilon$ and $wx_i + b - y_i \leq \epsilon$ for data points within / on the margin of tolerance that are above and below the hyperplane, respectively. Then, we introduce slack variables $\xi_i, \xi_i^* \geq 0$ representing the distance (error) between the margins of tolerance and data points that are outside the margins of tolerance above and below the hyperplane, respectively. Since the margin of tolerance is ignored, it only leaves us with the new slack variables:

$$min\frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4.13}$$

After applying the Lagrangian function and simplifying the formulated dual problem, we obtain:

$$w = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)x_i \tag{4.14}$$

Where λ_i and λ_i^* result from the Lagrangian formulation. Then, after applying Equation 4.14 to Equation 4.11 we get:

$$f(x) = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)x_i x + b \tag{4.15}$$

By defining the kernel function K, $K(x_i x) = \phi(x_i) \cdot \phi(x)$, with ϕ denoting the type of the kernel function (linear, polynomial, Gaussian, or sigmoid), we get:

$$f(x) = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)K(x_i x) + b \tag{4.16}$$

4.3.5 Deep neural networks

Artificial neural networks are modeled after the human brain's biological neural networks. They are extensively used in machine learning applications for different prediction tasks such as pattern recognition, classification, and regression [129].

An ANN is composed of three types of layers, namely: input layer, hidden layer(s), and output layer; a DNN is an ANN with two or more hidden layers [130]. The NN layers consist in a network of linked nodes known as neurons. Each neuron receives an input, processes it mathematically, and then produces a value. The inputs to the neurons in a layer come from the outputs of the neurons in the previous layer. This process goes on until the output layer is attained.

The neurons of the input layer (layer 0) are fed with the input data $x_i^{(0)}, 1 \leq i \leq N$. The rest of the layers' values (including the output layer) are defined as follows [131]:

$$x_i^{(k)} = f(\sum_{j=1}^{m} x_i^{(k-1)} w_{ij}^{(k)} + b_i) \tag{4.17}$$

The equation above represents the feedforward process of a neural network where the layers' nodes have each a value x, and a transfer function f. An activation function is what allows the neural network to learn complex relationships between dependent and independent variables through the non-linearities that are introduced. The nodes between layers are linked with connections which are characterized by the strength w that excites or inhibits nodes [129]. At the end of the feedforward process, the error value is measured through the loss function between the predicted value of the output layer and the actual desired output. The primary goal of a neural network is to optimize the error value and minimize it, in order to have a predicted output that is as close as possible to the desired output.

A crucial procedure for neural network training is backpropagation, which iteratively modifies the weights depending on the error measure obtained during forward propagation. The weights of the network are updated (see Equation 4.18) in a way that minimizes the error value E [132].

$$w_{ij}'^{(k)} = w_{ij}^{(k)} - \alpha \frac{\partial E}{\partial w_{ij}^{(k)}} \tag{4.18}$$

4.4 Atomic approach modeling

As mentioned above, atomic approach features descriptors are not uniform since each material has as many descriptors as its number of atoms, and materials do not have the same number of

atoms. Therefore, standard ML algorithms (including the previously mentioned ones) cannot be used on such data. To this end, a specific deep neural network topology was implemented for each atomic database to support atomic approach features descriptors. To achieve this, a modification was made in the feedforward process of the neural network. Given that an instance (S_i) from an atomic dataset is defined as the following, without considering the red variables:

$$S_i : \left\{ \begin{array}{lccc} d_1 = & [v_{11}, v_{12}, \cdots, v_{1n}] & \rightarrow & e_1 \\ d_2 = & [v_{21}, v_{22}, \cdots, v_{2n}] & \rightarrow & e_2 \\ \vdots & \vdots & \vdots & \vdots \\ d_m = & [v_{m1}, v_{m2}, \cdots, v_{mn}] & \rightarrow & e_m \end{array} \right\} \rightarrow E_i$$

A typical neural network should learn from the output E_i. In our case, the modification lays in the fact that our neural network does not learn from one output per structure but from as many outputs as there are atoms in a structure. In other words, for a particular structure, the neural network learns from each atom its atomic energy contribution to the total energy, which is referred to as $e_j (1 \leq j \leq m)$ in the representation above. The training process of such a neural network is described by the following algorithm:

For each structure S_i:

1. Create m (number of atoms) mini DNNs, each fed by its atomic descriptor $d_j (1 \leq j \leq m)$.
2. Predict the atomic energy contributions $e_j (1 \leq j \leq m)$ of all DNNs from step 1 using Equation 4.17.
3. Sum up all atomic energy contributions $e_j (1 \leq j \leq m)$ from step 2, and calculate the error $E(\sum_{j=1}^{m} e_j, E_i)$.
4. Update DNNs' weights using Equation 4.18, and repeat steps 2, 3, and 4 until the desired error is achieved.

Figure 4.2 illustrates the aforementioned algorithm.

4.5 Machine learning training process

The general process (see Figure 4.3) of machine learning prediction takes the features engineered dataset as an input for the selected models. The hyper-parameters of the models are

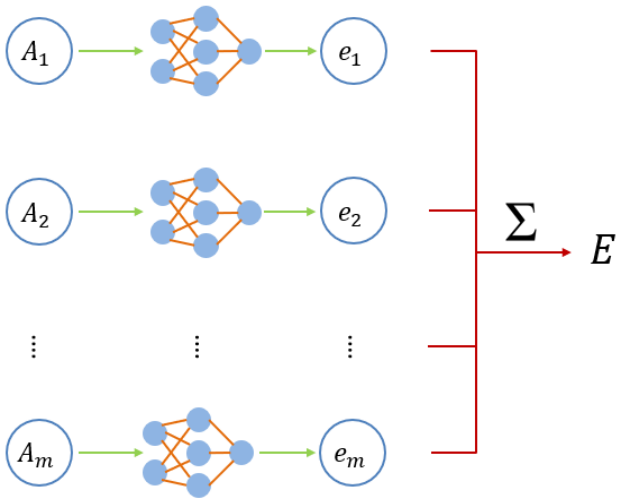


Figure 4.2: Proposed atomic deep neural network topology scheme for the energy property prediction.

set beforehand based on the best practices according to the state of the art. The models are then trained and evaluated through performance metrics. Depending on whether the models evaluation is satisfying or not, one would either proceed to the prediction process or return to the models settings in order to tune the hyper-parameters for better results.

In order to predict the energy property, the previously mentioned algorithms in this chapter were used as regression models. The general prediction formula is defined as follows:

$$E_i = RM_{hp}(DV_i) \tag{4.19}$$

Where:

1. E_i is the output energy value of structure i.

2. RM_{hp} is the regression model (EN, BR, RF, SVM, or DNN) built by learning the correlations between input data and target in the learning set. In order to optimize the investigated models and get the closest output to the target value, the hyper-parameters (HP) of the models were selected using a hyper-parameter tuning procedure, and out of all combinations, the configuration yielding the highest accuracy was chosen.

3. $DV_i = x_1, x_2, \cdots, x_n$ is the structure i's input descriptor vector of size n (60, 364 ,424, 60, 468, or 528) for the three structural/atomic datasets respectively generated using

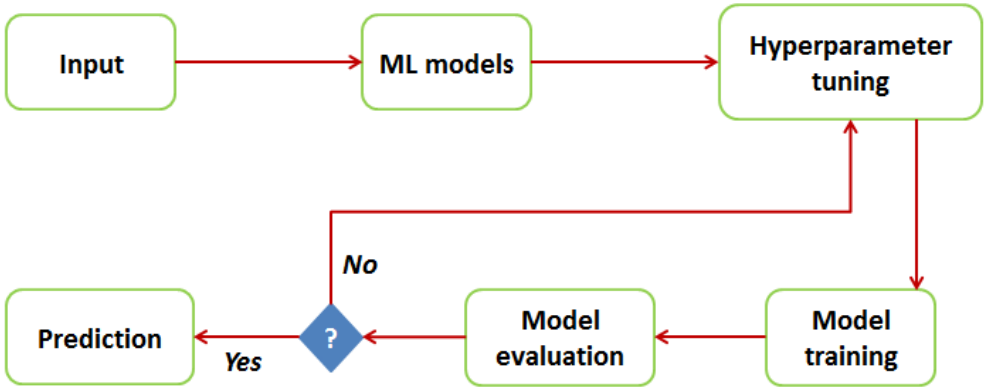


Figure 4.3: Flow chart of the training/testing process for the energy property prediction.

(2BDF-St, 3BDF-St, 2-3BDF-St, 2BDF-At, 3BDF-At, 2-3BDF-At) descriptors on the total eight merged raw databases.

Hyper-parameters are configuration options that are chosen before the model is trained and are not learnt from the data. They have substantial influence over the model's performance and have control over a number of learning-related variables. Therefore, hyper-parameter tuning is a crucial stage in machine learning since it may greatly enhance the model's performance and its capability for generalization. It allows to identify the ideal balance between model complexity and overfitting, thus producing better outcomes and more accurate predictions.

One of the most exhaustive hyper-parameter tuning strategies is the grid search. After defining a search space by specifying the potential values for each hyper-parameter, the grid search strategy consists in defining all possible combinations of hyper-parameters and train the ML model with each combination. Then, the hyper-parameters combination yielding the best performance is selected to build the model. While this strategy is very efficient since it explores every possible combination, it would most probably be computationally out of reach [133].

One could reduce the computational time and cost by using the random search strategy instead of the grid search one [133]. Although this strategy defines all possible combinations of hyper-parameters as well, it only trains the ML model with randomly selected subsets of the hyper-parameter combinations. While random search is less computationally costly, it is definitely less efficient since it does not cover all possible hyper-parameter combinations and thus, it may never reach the best performing hyper-parameter combination.

For this purpose, another method was used for hyper-parameter tuning in this study. It is called grid search cross validation [134]; it allows one to fully benefit from the grid search strategy while reducing the computational cost. This is achieved by dividing the input dataset and the hyper-parameter combination set into a number (k) of folds. Then, a hyper-parameter combination subset is attributed to each input fold for the model training. Finally, the hyper-parameter combination subset attributed to the best performing fold is selected.

In this study, a total of 18 (5 previously mentioned structural approach ML algorithms $\times$ 3 structural approach's features descriptors + 1 atomic approach ML algorithm $\times$ 3 atomic approach's features descriptors) models were implemented to predict the energy value. Their architecture and configuration respectively depend on the features descriptor size and hyper-parameter tuning outcome. Table 4.1 summarizes the investigated ML models along with their respective configurations.

In this study, the different investigated algorithms are tuned as follows:

- ElasticNet: to fully benefit from L1 and L2 penalties, the λ value was set to 1; also, the alpha hyper-parameter was set 0.5 in order to equally take advantage from both penalties.

- Bayesian ridge: the α_1 and α_2 parameters respectively representing the prior distribution precision for the weights and the noise were set to small values for a stronger regularization.

- Random forest: no pruning was performed in the random forest training to let the trees grow to their maximum depth which allows to capture tricky and complex relationships in the data. The selected criterion was the MSE loss and the bootstrap Boolean parameter was set to "True" to allows to introduce diversity and randomness.

- Support vector machine: the "C" parameter controlling tradeoff between prediction accuracy and cost was set to 1, the epsilon tolerance margin to 0.01, and the chosen kernel is RBF for its strong ability to model non-linearities.

- Deep neural network: For the DNN models' architecture (of both the structural and atomic approaches), the hidden layers number is comprised between 2 and 6 (depending on the structural and atomic descriptors used); the transfer function (TF) used in the input and hidden layers is ReLU which has proven to be very effective in deep learning [135]. As for the output layer, the Linear Activation function, most suitable transfer

Approach	Model	Features	Description	Hyper-parameters
Structural	EN	2BDF	In= 60,Out= 1	$\lambda = 1$
		3BDF	In= 364,Out= 1	$\text{L1}_{ratio} = \alpha = 0.5$
		2-3BDF	In= 424,Out= 1	Selection = cyclic
	BR	2BDF	In= 60,Out= 1	$\alpha_1 = \alpha_2 = \alpha_3 = 1e^{-6}$
		3BDF	In= 364,Out= 1	$\lambda_1 = \lambda_2 = \lambda_3 = 1e^{-6}$
		2-3BDF	In= 424,Out= 1	
	RF	2BDF	In= 60,Out= 1	$\text{Ccp}_{alpha} = 0$
		3BDF	In= 364,Out= 1	criterion = MSE
		2-3BDF	In= 424,Out= 1	bootstrap = True
	SVM	2BDF	In= 60, Out= 1	C = 1
		3BDF	In= 364, Out= 1	epsilon = 0.01
		2-3BDF	In= 424, Out= 1	Kernel = RBF
	DNN	2BDF	In= 60,HL= 2, Out= 1	TF: ReLU, Linear
		3BDF	In= 364, HL= 4, Out= 1	LA: Adam
		2-3BDF	In= 424, HL = 5, Out= 1	LF: MSE
Atomic	DNN	2BDF	In= 60, HL= 2, Out = 1	TF: ReLU, Linear
		3BDF	In= 468, HL= 5, Out= 1	LA: Adam
		2-3BDF	In= 528, HL= 6, Out= 1	LF: MSE

Table 4.1: Machine learning models investigated in this study with their respective hyper-parameters.

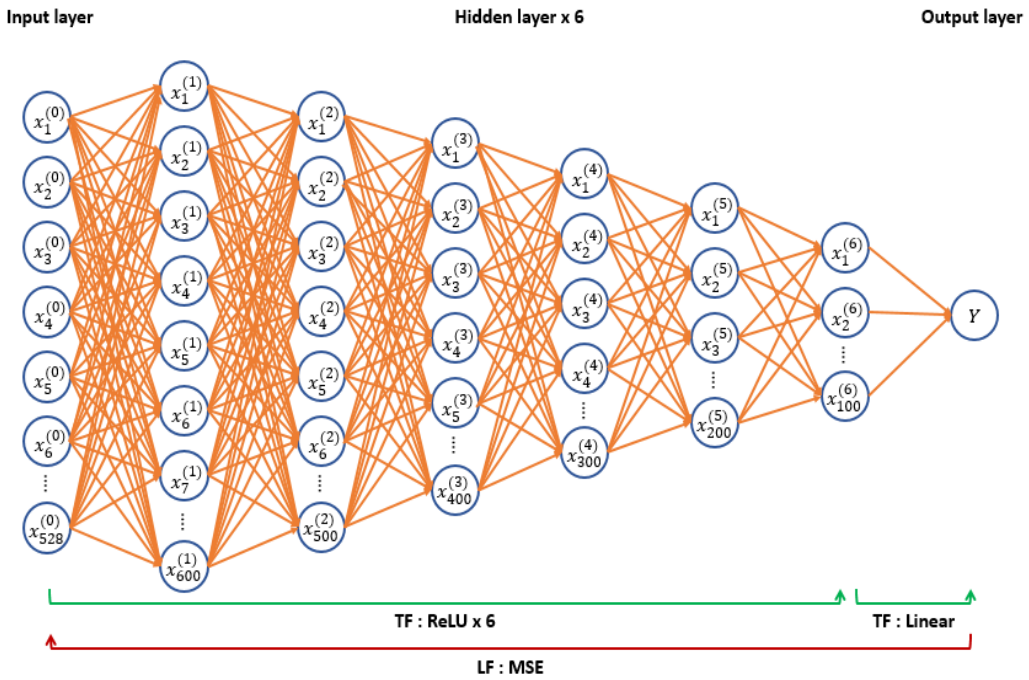


Figure 4.4: Schematic representation of the 2-3BDF atomic DNN architecture.

function for regression problems [136], was used. The learning algorithm (LA) and the loss function (LF) implemented were Adam optimizer and MSE, respectively. To provide a better understanding of the proposed model, figure 4.4 represents a schematic representation of the atomic DNN architecture. For a comprehensive overview, the illustrated model corresponds to that of the 2-3BDF-At DNN.

4.6 Conclusion

Machine learning has the fascinating ability to discern and identify patterns in data. Given the availability of engineered crystallographic data represented with structural and atomic descriptors, we began by analyzing the fundamental components of predictive models, including approaches, strategies, and algorithms. The strength and power of the latter were exploited to construct models for the energy property prediction.

To this end, the structural approach consisted of five machine / deep learning models implementation for the prediction task, to which the three structural descriptors were fed as an input. Moreover, an unconventional neural network topology was proposed and implemented to support atomic descriptors for the prediction of energy. In the next chapter,

the evaluation of the several proposed implementations will be performed and the obtained results will be presented.

Chapter 5

Results and Discussion of Crystal Structure Energy Prediction

"There is a magic in graphs. The profile of a curve reveals in a flash a whole situation." – Henry D. Hubbard.

5.1 Introduction

Machine learning models are evaluated using statistical measures with the sole purpose of assessing the effectiveness of a model's predictive ability. These metrics offer measurable indicators of a model's performance. They help one to fairly evaluate many models, choosing the one that performs the best, and pinpointing potential areas for improvement. The assessment of machine learning predictions is an essential step in the complex world of crystal structure prediction, as the predictive ability and efficiency of the ML models are carefully evaluated. In our case, we focus on predicting the energy property of crystal structures which is a fundamental pursuit in materials science. The results in this chapter summarize not only the complexity of our crystal structure energy prediction estimates but also the need of thorough analysis in the field of predictive modeling.

As we move on with this chapter, we will proceed to the evaluation of our implemented models through well-chosen performance metrics. Our objective is twofold: to select and validate the strongest crystal structure descriptor, and to identify the best energy prediction ML model. This evaluation phase is crucial in order to determine the strengths and drawbacks of the implemented models and to guide the future course of the crystal structure energy prediction investigation.

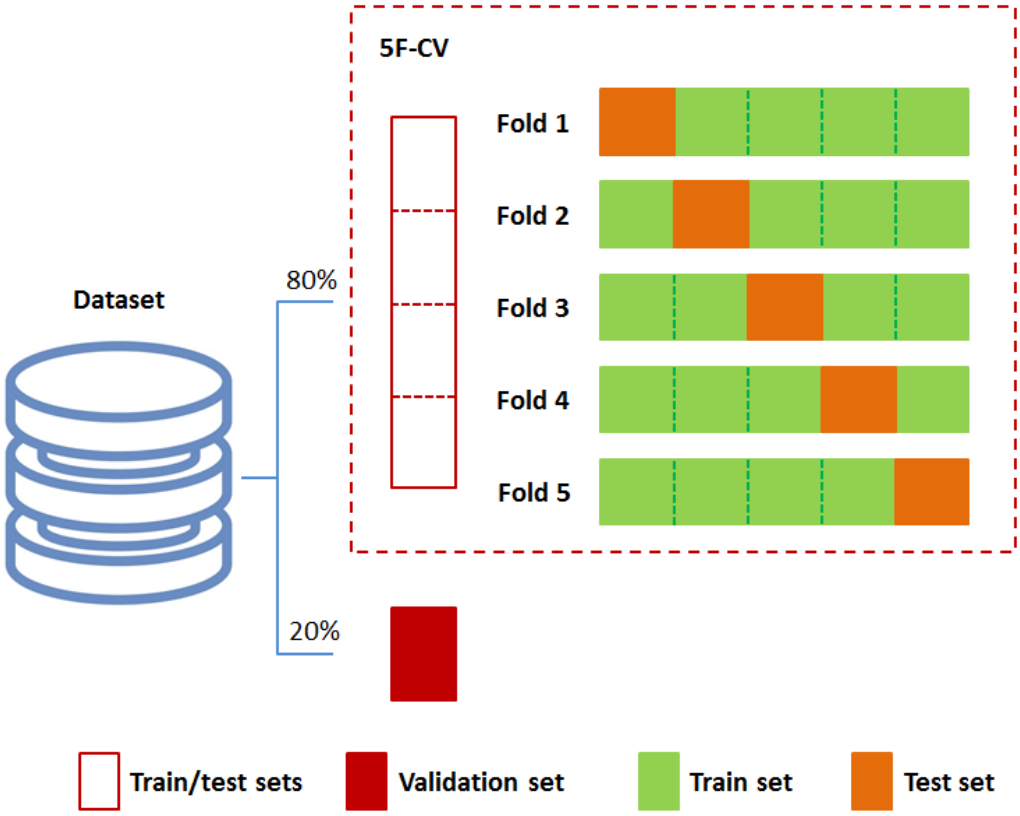


Figure 5.1: Depiction of the data split procedure into 80% training/testing (with 5-fold cross-validation process) and 20% validation sets.

5.2 Evaluation strategy

In this study, we aim at predicting the energy property of crystal structures. Several ML algorithms were selected to model the relationship between the features descriptors extracted from the structure/atom-based approaches and the energy property. In order to evaluate and validate the prediction process, an evaluation strategy was adopted as explained hereafter.

5.2.1 Data split

For both the structural and atomic approaches, the six datasets resulting from the features extraction process were each shuffled and divided into two subgroups as depicted in Figure 5.1.

- The first subgroup comprising 80% of the dataset is identified as the training and testing

set. The performance of this subgroup was examined using five-fold cross validation (5F-CV) procedure, dividing the dataset into five mutually selected train/test partitions of equal size. For each of the five test runs, one partition was assigned as the test set and the remaining samples were used to train the models.

- The second subgroup of the 20% remaining data was used for the final validation stage.

This technique helps one gain insights into the ML model's generalizability and performance capabilities. In fact, incorporating k-fold cross-validation technique allows us to assess the extent of a model's robustness, thus, making the overall prediction performance more comprehensive to examine and evaluate.

5.2.2 Evaluation metrics

The evaluation procedure of the different investigated algorithms was performed using two kind of metrics, namely: a graphical-based assessment and a statistical-based assessment.

The graphical-based assessment is represented through the Receiver Operating Characteristics (ROC) and the Area Under Curve index. ROC/AUC is indeed an effective metric to evaluate the algorithm's learning ability [137]. In short, ROC provides a systematic analysis through a graph which expresses the balance between benefits and costs [138] and displays threshold between the sensitivity (Sn(e,t) placed across the abscissa axis) and 1-specificity (Sp(e,t) plotted along the ordinate axis) [139] provided by [140]:

$$Sn(e,t) = \frac{\sum TP(e,t)}{\sum TP(e,t) + \sum FN(e,t)} \tag{5.1}$$

$$Sp(e,t) = \frac{\sum TN(e,t)}{\sum TN(e,t) + \sum FP(e,t)} \tag{5.2}$$

With TP, FN, TN, and FP representing True Positive, False Negative, True Negative, and False Positive, respectively.

The AUC is a measure used to analyze the efficiency of the algorithm; its value is bounded between 0 and 1 and represents the area under ROC curve of $Sn(e,t)$ or true positive rate (TPR) versus $Sp(e,t)$ or false positive rate (FPR) [141]. The higher AUC score, the better the performance.

$$AUC = \int_0^1 TPRd(FPR) \tag{5.3}$$

As for the statistical-based assessment, measures such as mean squared error, mean absolute error, and R^2 were considered. MSE and MAE are two representations of the difference (error) between the predicted / expected ($\hat{y}_i$) value and the true / observed (y_i) value for an instance; they are usually used in regression problems and are defined as follows [142]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.5}$$

R^2, also known as coefficient of determination, represents how well a regression model's fitness is. Through a scale of 1 to 100%, it measures the strength of the relationship between the dependent variables and the model; a higher percentage of R^2 means a better performance of the model [143]. The equation used to calculate R^2 is [144]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5.6}$$

These metrics were not only used to evaluate the algorithms ability to accurately predict the energy property, but also to compare the developed alternative models in order to identify the best performance and, consequently, the best input features from the aforementioned descriptors for energy prediction.

5.3 Experimental setup

In this study, we aim to conduct an investigation of crystal structure energy predication. For this purpose, a computational database was generated using USPEX code at Oganov's Lab which was used for the training procedure to construct machine learning models. The modeling stage was performed with Python 3 programming language through the frameworks Keras-Tensorflow [145] and Pytorch [146] for the structural and atomic approaches respectively using GPU execution mode. Scikit-learn [147] library was used for metrics calculation to evaluate the models' performance, and results plots were generated using Matplotlib [148].

5.4 Structural approach results interpretation

This section presents the ML investigation and performance evaluation through an assessment based on the previously mentioned metrics. The developed structural-based models will be

analyzed and compared to yield an efficient model for features validation and energy property prediction. The results of the structural approach will be presented separately according to the utilized descriptor (2BDF-St, 3BDF-St, and 2-3BDF-St). For each descriptor, the performance of the learning models will be illustrated through a fold-based and a model-based comparison visualization.

When analyzing and examining the fold-based and model-based performance results, we should respectively consider the following two main factors:

- **Balanced data distribution.** In order to perform a 5-fold cross-validation, data extracted from the different descriptors is evenly divided into 5 subgroups, each representing a fold. To make sure that the data distribution is balanced between folds with respect to data diversity in terms of features, the performance of the folds should not reveal a significant difference from one another. This is important insofar as the good performance of a model should not depend solely on the distribution of a certain fold. Indeed, the model must be able to yield a fair prediction result regardless of the fold, which insures the model generalizability.

- **Metrics harmony.** The evaluation of the investigated ML models is carried through performance metrics including MSE, MAE, R^2, and ROC/AUC. In the assessment process, no discrepancy or inconsistency should occur in the performance metrics values. This means that, in the fold-based comparison, if a certain fold yields a lower MSE, it should also yield a lower MAE and a higher R^2 and AUC values. Likewise, in the model-based comparison, the metrics should present with the same harmony and consistency.

5.4.1 Prediction results of 2BDF-St-based models

2BDF-St is a structural-based descriptor using the two-body distribution function representation. It is defined by a 60-element vector for each structure. We present herein the prediction results of the energy property of crystal structures represented through this descriptor.

5.4.1.1 Fold-based results comparison

Figure 5.2 displays the test phase prediction performance of the 2BDF-St models (EN, BR, RF, SVM, and DNN) with regards to MSE, MAE, and R^2 metrics.

Each metric of every subplot of Figure 5.2 is represented by 5 bars (one for each fold), where the best performing fold is framed in a black edge color. The three metrics were plotted

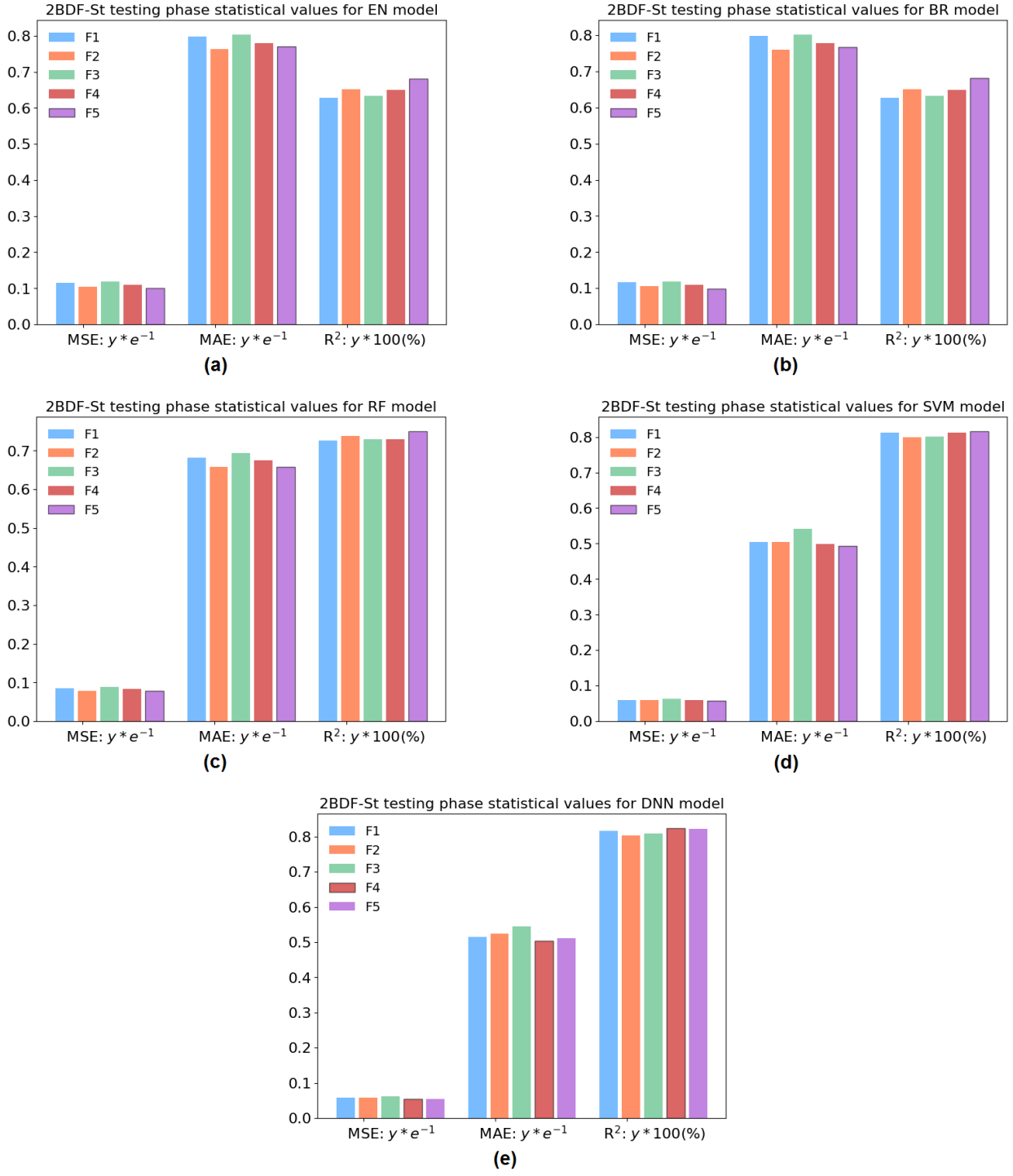


Figure 5.2: Bar plots of energy prediction's testing phase of 2BDF-St models folds performance with regards to MSE, MAE, and R^2 which values are obtained as ye^{-1}, ye^{-1}, and $y \times 100\%$, respectively.

using the same scale and their values as mentioned in the subplots' x axis are obtained as such: ye^{-1}, ye^{-1}, and $y \times 100$ (as a percentage value), with y being the statistical value on the y axis for MSE, MAE, and R^2, respectively.

The analysis of the bar plots depicting the investigated models performance reveals a remarkable consistency in the results. Indeed, each fold's performance of every ML model stays closely clustered, indicating an almost equivalent predictive ability across the various dataset subsets. We also observe that Fold 5 has a little edge over the others yielding better MSE, MAE, and R^2 values for the models EN, BR, RF, and SVM. Moreover, Fold 4 exhibits a marginal advantage over, not only, the other folds of the DNN model but also all the other folds of all models combined.

For more details, Table 5.1 presents precise MSE, MAE, and R^2 values for every model's five folds prediction performance.

In Table 5.1, we find, for each 2BDF-St models folds, values that are bold-emphasized. These values represent the best results of each model (lowest MSE, lowest MAE, and higher R^2). In addition, the best performing fold of all models combined is indicated in underlined values.

It is clearly noted that the difference between MSE, MAE, and R^2 values of every model's folds is small; this ensures that the data distribution is well balanced between all five folds of the 2BDF-St descriptor's dataset.

5.4.1.2 Model-based results comparison

In order to compare the best performing folds from each 2BDF-St model, we present Figure 5.3.

In Sub-figure 5.3 (a), the MSE, MAE, and R^2 values of the (best performing) models EN-St-F5, BR-St-F5, RF-St-F5, SVM-St-F5, and DNN-St-F4 are plotted together for a better comparison visibility. Since the statistical metrics MSE, MAE, and R^2 are plotted on the same graph using the same scale (from 0 to 1) on the y axis, their values as mentioned in the subplots' legend are obtained respectively as follows: ye^{-1}, ye^{-1}, and $y \times 100$ (as a percentage value), with y being the statistical value on the y axis. Sub-figure 5.3 (b), on the other hand, illustrates the ROC curve plot of the same previously mentioned models.

We note that the plots of Figure 5.3 exhibit a coherence between the different quality metrics. Indeed, where MSE and MAE metrics are lower, R^2 and AUC values are higher. This ensures the harmony between these performance metrics which is essential for results credibility.

Model	Fold	MSE	MAE	R^2
EN	Fold 1	0.0116	0.0798	62.76
	Fold 2	0.0105	0.0763	65.18
	Fold 3	0.0119	0.0803	63.28
	Fold 4	0.0109	0.0780	65.01
	Fold 5	**0.0100**	**0.0771**	**68.01**
BR	Fold 1	0.0116	0.0798	62.73
	Fold 2	0.0105	0.0791	65.16
	Fold 3	0.0119	0.0802	63.22
	Fold 4	0.0109	0.0778	64.87
	Fold 5	**0.0099**	**0.0767**	**68.14**
RF	Fold 1	0.0085	0.0682	72.67
	Fold 2	0.0079	0.0659	73.82
	Fold 3	0.0088	0.0694	72.91
	Fold 4	0.0084	0.0675	72.99
	Fold 5	**0.0078**	**0.0659**	**75.01**
SVM	Fold 1	0.0059	0.0504	81.14
	Fold 2	0.0060	0.0504	80.00
	Fold 3	0.0064	0.0541	80.17
	Fold 4	0.0059	0.0499	81.17
	Fold 5	**0.0057**	**0.0494**	**81.2**0
DNN	Fold 1	0.0058	0.0515	81.55
	Fold 2	0.0059	0.0525	80.24
	Fold 3	0.0062	0.0545	80.93
	Fold 4	**0.0055**	**0.0503**	**82.36**
	Fold 5	0.0055	0.0511	82.20

Table 5.1: Numeric results of MSE, MAE, and R^2 quality metrics assessing ML performance used with 2BDF-St features descriptor.

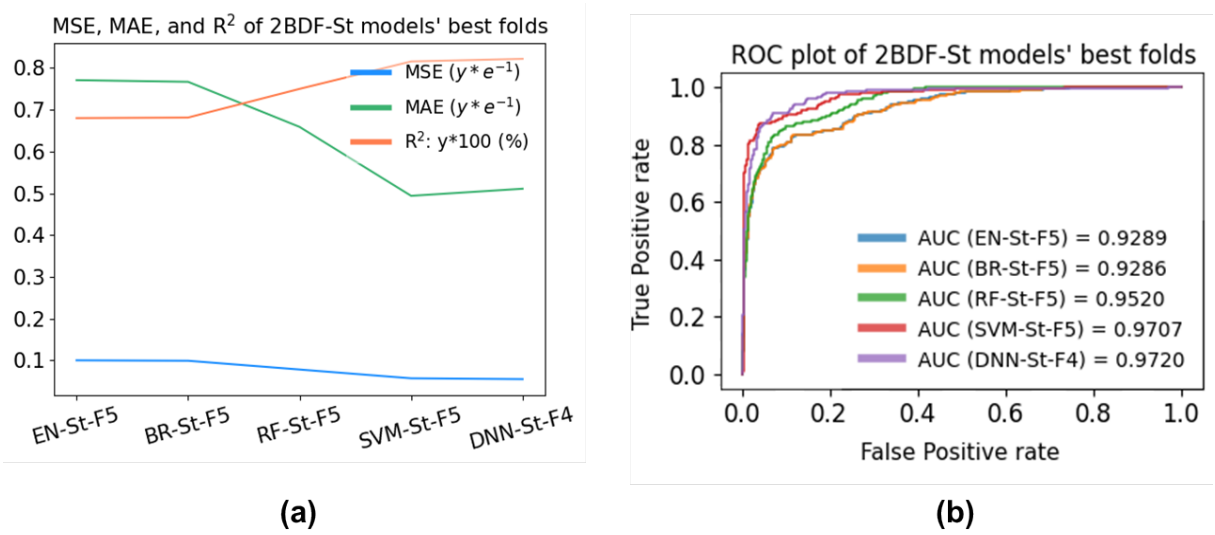


Figure 5.3: Accuracy assessment of energy prediction's testing phase of the 2BDF-St models' best performing fold. (a) MSE, MAE, R^2 measures, (b) ROC plot.

The best selected folds of each investigated ML model were all able to perform the energy prediction using the 2BDF-St descriptors. They all lead to satisfying results where SVM-St-F5 and DNN-St-F4 had an edge over other models while EN-St-F5 and BR-St-F5 yielded the least favorable results and RF-St-F5 returned an average outcome. The best recorded performance is that of DNN-St-F4 with a score of 0.0055, 0.0503, 82.36%, and 0.9720 for MSE, MAE, R^2, and AUC, respectively.

5.4.2 Prediction results of 3BDF-St-based models

In this section, the energy prediction results, as related to the structural three-body distribution function descriptor, will be analyzed. The features of this descriptor are represented through a vector of size 364. The models EN, BR, FR, SVM, and DNN will be trained and tested using 5-fold cross-validation technique for the task of energy prediction.

5.4.2.1 Fold-based results comparison

In order to analyze and examine the performance of the investigated ML models using 3BDF-St in a fold-wise manner, we present the testing phase prediction results of every fold from each model in Figure 5.4.

The bars of each subplot in Figure 5.4 are gathered into 3 groups of 5 bars, with 3 being the number of metrics MSE, MAE, and R^2 (for which the value is respectively obtained as

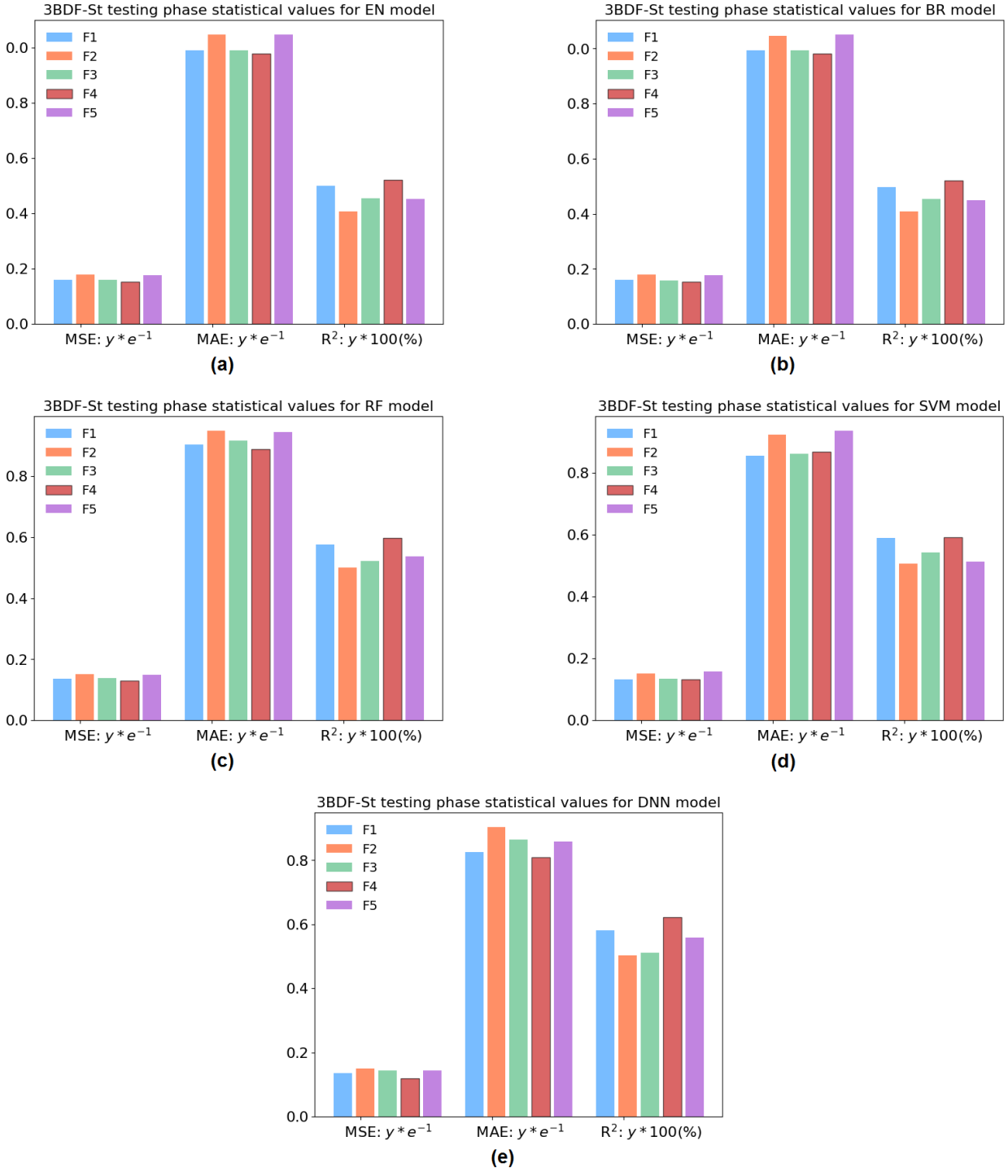


Figure 5.4: Bar plots of energy prediction's testing phase of 3BDF-St models folds performance with regards to MSE, MAE, and R^2 which values are obtained as ye^{-1}, ye^{-1}, and $y \times 100\%$, respectively.

ye^{-1}, ye^{-1}, and $y \times 100\%$) and 5 the number of folds.

Upon the examination of the bar chart featuring the models' testing results with regards to MSE, MAE, and R^2, a noteworthy pattern emerges. We observe a clear uniformity across the performance of our cross-validation procedure's five distinct folds. This uniformity indicates a consistent ability to predict across different data subsets. Although all folds yielded good results, which underlines the investigated models' robustness in addressing the energy prediction challenge, it is essential to emphasize that Fold 4 demonstrates a modest performance advantage over the others, showing marginally better outcomes in terms of MSE, MAE, and R^2.

Table 5.2 provides accurate MSE, MAE, and R^2 values for the five-fold prediction performance of each model.

In Table 5.2, for each examined model, the best performing fold in terms of the selected quality metrics is emphasized in bold. Moreover, as indicated in underlined values, the 4^{th} fold of the DNN-St model outperformed all other folds including those of the remaining models.

In the 3BDF-St descriptor's dataset, the data is proven to be evenly distributed across all five folds through the minimal disparities in MSE, MAE, and R^2 values of Table 5.2.

5.4.2.2 Model-based results comparison

In the purpose of highlighting the model-based performance comparison, we introduce Figure 5.5.

The subplots (a) and (b) of Figure 5.5 respectively illustrate a visual representation of the statistical and graphical results of the best performing fold of each investigated ML model. We observe that Fold 4 outperformed the other folds within every ML model performance; however, the comparison analysis revealed that this fold's data distribution yielded distinct results from one model to another.

The statistical metrics results of EN-St-F4, BR-St-F4, RF-St-F4, SVM-St-F4, and DNN-St-F4 plotted in Figure 5.5 (a) are obtained using the scale ye^{-1}, ye^{-1}, and$y \times 100(\%)$ for MSE, MAE, and R^2. This plot shows a consistent behavior with that of the ROC plot of Figure 5.5 (b). We notice that the lowest MSE and MAE values were recorded for DNN-St-F4 with a score of 0.0120 and 0.0809 respectively. This same model yielded the highest R^2 and AUC score of 62.24% and 0.9249 respectively. The RF-St-F4 and SVM-St-F4 models yielded approximate average results while the remaining EN-St-F4 and BR-St-F4 models recorded the highest MSE and MAE values as well as the lowest R^2 and AUC ones, making them

Model	Fold	MSE	MAE	R^2
EN	Fold 1	0.0159	0.0990	50.05
	Fold 2	0.0179	0.1048	40.76
	Fold 3	0.0159	0.0992	45.61
	Fold 4	**0.0153**	**0.0979**	**52.19**
	Fold 5	0.0177	0.1048	45.25
BR	Fold 1	0.0160	0.0993	49.77
	Fold 2	0.0179	0.1047	40.89
	Fold 3	0.159	0.0994	45.47
	Fold 4	**0.0153**	**0.0981**	**52.05**
	Fold 5	0.0178	0.1051	4501
RF	Fold 1	0.0135	0.0903	57.70
	Fold 2	0.0151	0.0949	50.12
	Fold 3	0.0139	0.0916	52.31
	Fold 4	**0.0129**	**0.0888**	**59.70**
	Fold 5	0.0149	0.0944	53.71
SVM	Fold 1	**0.0131**	**0.0855**	58.92
	Fold 2	0.0150	0.0922	50.53
	Fold 3	0.0134	0.0860	54.10
	Fold 4	**0.0131**	0.0868	**59.03**
	Fold 5	0.0158	0.0935	51.14
DNN	Fold 1	0.0136	0.0825	58.03
	Fold 2	0.0150	0.0903	50.35
	Fold 3	0.0143	0.0865	51.04
	Fold 4	**0.0120**	**0.0809**	**62.24**
	Fold 5	0.0143	0.0859	55.77

Table 5.2: Numeric results of MSE, MAE, and R^2 quality metrics assessing ML performance used with 3BDF-St features descriptor.

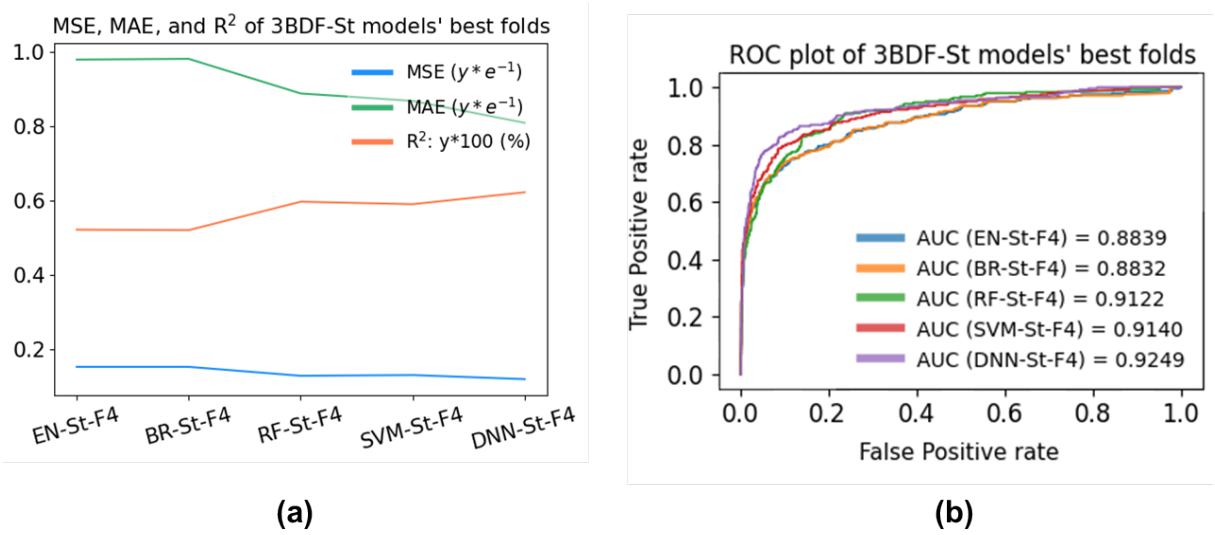


Figure 5.5: Accuracy assessment of energy prediction's testing phase of the 3BDF-St models' best performing fold. (a) MSE, MAE, R^2 measures, (b) ROC plot.

the worst performing models in terms of the energy property prediction for the 3BDF-St descriptor dataset.

The kinship between the different quality metrics results for the 3BDF-St descriptor dataset shows no discrepancy, thus proving their harmonic behavior.

5.4.3 Prediction results of 2-3BDF-St-based models

The third dataset of the structural approach for crystal structure's energy prediction is the combination the two previously mentioned datasets (2BDF-St and 3BDF-St) which we refer to as 2-3BDF-St. This dataset is generated using both two- and three-body distribution functions and is represented through a vector of 424 elements.

The next two subsections present the energy prediction results of the selected ML models performance using the descriptor 2-3BDF-St in terms of fold- and model-based analysis.

5.4.3.1 Fold-based results comparison

The structural approach's five ML models were each trained and tested on five folds subsets of the 2-3BDF-St's descriptor dataset. The testing performance of these models' folds with regards to the selected quality metrics is illustrated in Figure 5.6.

Figure 5.6 shows the testing phase performance results of the several implemented ML models with the structural descriptor 2-3BDF-St. The results are depicted in terms of the

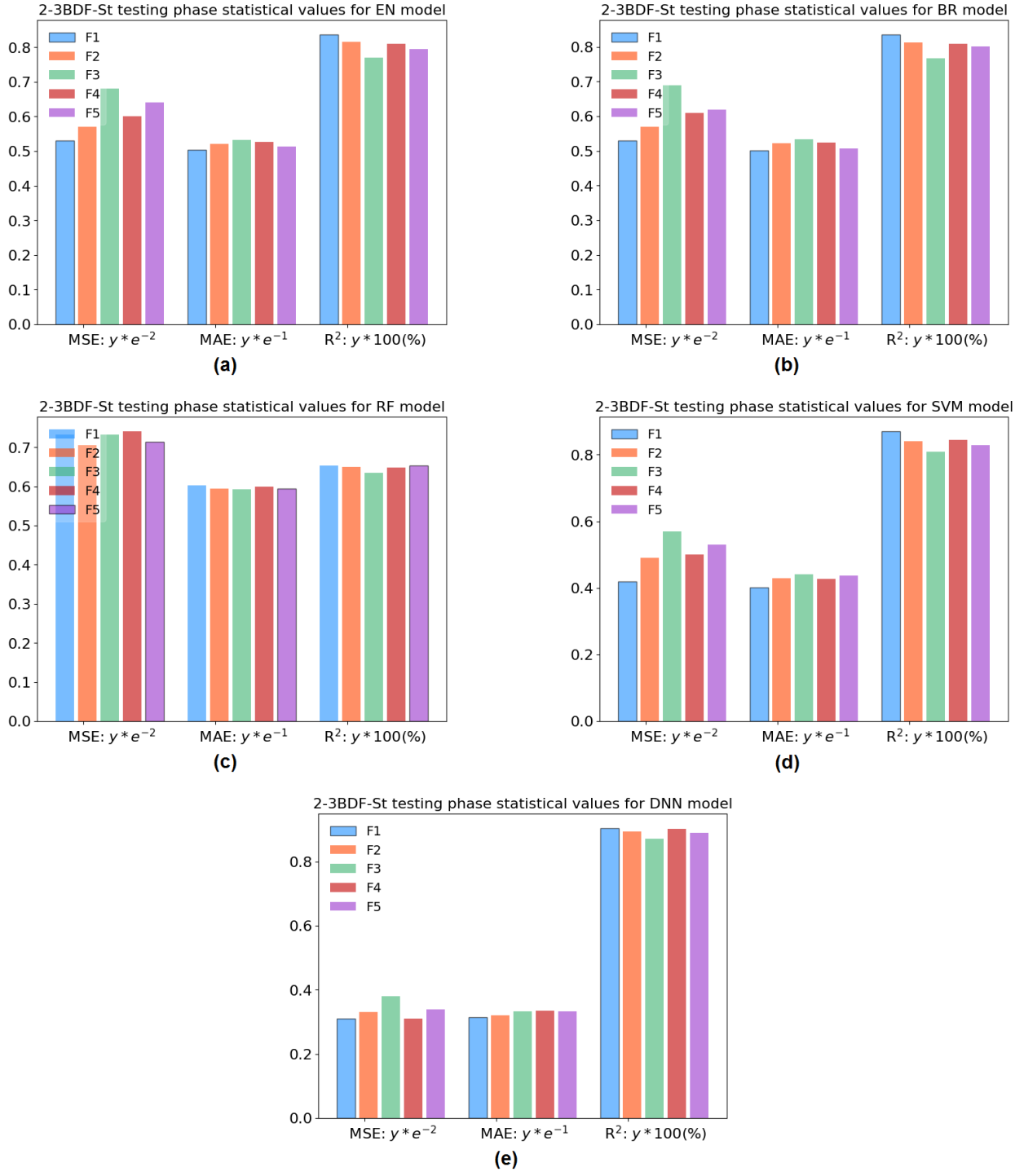


Figure 5.6: Bar plots of energy prediction's testing phase of 2-3BDF-St models folds performance with regards to MSE, MAE, and R^2 which values are obtained as ye^{-2}, ye^{-1}, and $y \times 100\%$, respectively.

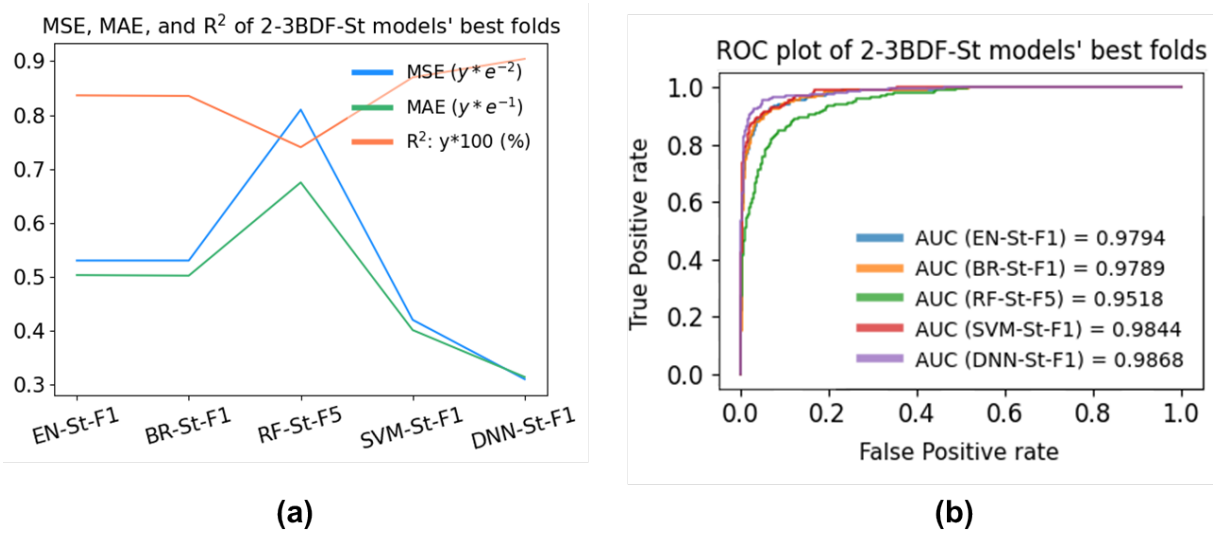


Figure 5.7: Accuracy assessment of energy prediction's testing phase of the 2-3BDF-St models' best performing fold. (a) MSE, MAE, R^2 measures, (b) ROC plot.

quality metrics MSE, MAE, and R^2 which are obtained from the bar plots as follows: ye^{-2}, ye^{-1}, and $y \times 100\%$, respectively.

The examination of bar charts in Figure 5.6 displays that the 2-3BDF-St dataset's five folds are undoubtedly well balanced in terms of data distribution. All five folds yielded, to a certain extent, approximate prediction results. As highlighted, the bars with a black edge indicate the best performing fold. We notice that Fold 1 surpassed the other ones for the models EN, BR, SVM, and DNN, while the RF implementation showed that the outperforming fold was the fifth one.

Detailed quality metrics values of the five models' testing performance are summarized in Table 5.3.

As mentioned above, the overall energy prediction performance as related to the 2-3BDF-St descriptor dataset revealed that FoldS 1 and 5 (highlighted in bold in Table 5.3) had a slight advantage over the three remaining folds. Among all models, DNN-St's 1^{st} fold yielded the best predictive ability of the energy property (as underlined in Table 5.3).

5.4.3.2 Model-based results comparison

After selecting the best performing fold of each investigated model implemented with the 2-3BDF-St descriptor, we perform a model-based comparison analysis of these folds as shown in Figure 5.7.

Figure 5.7 (a) represents the statistical quality metrics MSE, MAE, and R^2 (with the scale

Model	Fold	MSE	MAE	R^2
EN	**Fold 1**	**0.0053**	**0.0503**	**83.65**
	Fold 2	0.0057	0.0521	81.57
	Fold 3	0.0068	0.0533	77.03
	Fold 4	0.0060	0.0527	81.06
	Fold 5	0.0064	0.0513	79.50
BR	**Fold 1**	**0.0053**	**0.0502**	**83.54**
	Fold 2	0.0057	0.0523	81.35
	Fold 3	0.0069	0.0533	76.81
	Fold 4	0.0061	0.0524	80.90
	Fold 5	0.0062	0.0507	80.08
RF	Fold 1	0.0083	0.0683	74.07
	Fold 2	0.0080	0.0675	73.78
	Fold 3	0.0083	0.0673	71.90
	Fold 4	0.0084	0.0680	73.60
	Fold 5	**0.0081**	**0.0675**	**74.03**
SVM	**Fold 1**	**0.0042**	**0.0401**	**86.99**
	Fold 2	0.0049	0.0430	83.95
	Fold 3	0.0057	0.0441	80.84
	Fold 4	0.0050	0.0427	84.38
	Fold 5	0.0053	0.0437	82.93
DNN	**Fold 1**	**0.0031**	**0.0314**	**90.42**
	Fold 2	0.0033	0.0321	89.41
	Fold 3	0.0038	0.0333	87.05
	Fold 4	0.0031	0.0335	90.27
	Fold 5	0.0034	0.0333	88.97

Table 5.3: Numeric results of MSE, MAE, and R^2 quality metrics assessing ML performance used with 2-3BDF-St features descriptor.

ye^{-2}, ye^{-1}, and $y \times 100\%$, respectively) obtained by the best performing folds for each ML model with the 2-3BDF-St descriptor. The comparison analysis shows that the deep neural network model exceeded the other models in terms of energy predictive ability with a score of 0.0031, 0.0314, and 90.42 for MSE, MAE, and R^2, respectively. This leading performance was followed by that of the support vector machine model. The ElasticNet and Bayesian ridge models yielded average results, while the random forest model achieved the least satisfactory outcome.

The acquired statistical quality metrics results pattern is backed up by Figure 5.7 (b) representing the graphical ROC plot of the same models. The best AUC value score is that of the DNN-St-F1 model (0.9868) followed by SVM-St-F1, EN-St-F1, BR-St-F1, and RF-St-F5 in this order.

To summarize, we observed a similar behavior for all three structural-based models. Indeed, deep neural networks outperformed the other 4 models, followed by SVM as the second-best performing model in all three cases, while the lowest results were reported by both EN and BR for 2BDF-St and 3-BDF-St with a similar performance. However, RF achieved average results for 2BDF-St and 3BDF-St descriptors, but its predictive ability significantly dropped with the 2-3BDF-St descriptor due to the fact that it cannot handle bigger datasets (in terms of features number).

In a descriptor-wise analysis, the obtained results show that all the algorithms produced promising results with the 2BDF-St database, while the 3BDF-St database led to a decrease in the performance of the algorithms. However, important results were obtained with the 2-3BDF-St database, with the notable exception of RF, which performed better using 2BDF-St database.

5.5 Atomic approach results interpretation

This section presents the performance evaluation of the atomic-based deep neural network developed through an assessment based on the previously mentioned metrics. The developed atomic-based models will be analyzed and compared to yield an efficient model for features validation and energy property prediction.

As opposed to the structural approach, in this atomic approach we only dispose of one model that is a deep neural network. This decision was made based on the fact that atomic descriptors are not uniform and are unfit for any machine learning process. For this purpose, an unconventional deep neural network topology was developed to support these descriptors.

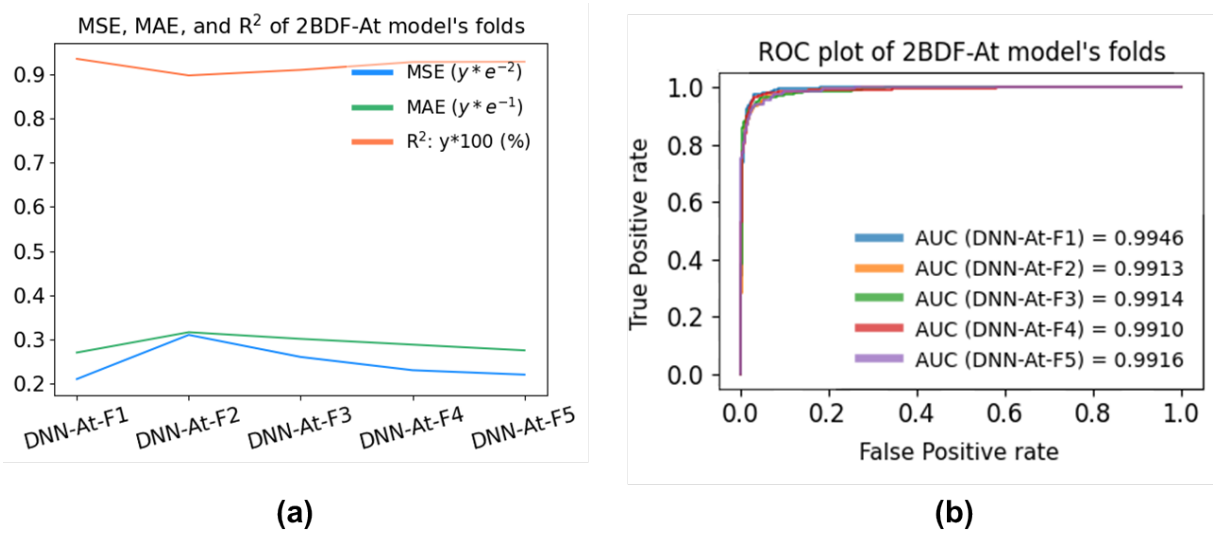


Figure 5.8: Accuracy assessment of energy prediction's testing phase of the 2BDF-At models' 5-folds. (a) MSE, MAE, R^2 measures, (b) ROC plot.

Therefore, in the following subsections, we will proceed with a fold-based comparative analysis for the energy property prediction.

5.5.1 Prediction results of 2BDF-At-based models

In this section, the results related to the energy prediction using 2BDF-At descriptor are presented. 2BDF-At is a dataset generated using the two-body distribution function with the atomic approach. This descriptor is an atom-wise representation of the input data. It is defined by a structure composed of n_i vectors of size 60, with n_i being the number of atoms of the i^{th} crystal structure in the dataset.

Figure 5.8 depicts the statistical and graphical results of the energy prediction produced by the 2BDF-At's 5 folds of the cross-validation process.

The analysis of the statistical metrics illustrated in Figure 5.8 (a) as well as the ROC plot in Figure 5.8 (b) shows that there is a narrow difference between performance results achieved by the 5 folds of the 2BDF-At model in the testing phase. We also notice that there is no observable inconsistency when it comes to the patterns of the plotted quality metrics, thus, ensuring the reliability of the results.

We present in Table 5.4 the numeric values of the testing phase energy prediction of each 2BDF-At DNN's folds.

The examination of Table 5.4 demonstrates that all five folds of the 2BDF-At model yielded a strong performance and resulted a promising outcome. With only slight differences

Model	MSE	MAE	R^2	AUC
DNN-At-F1	**0.0021**	**0.0270**	**93.51**	**0.9946**
DNN-At -F2	0.0031	0.0316	89.76	0.9913
DNN-At -F3	0.0026	0.0301	91.01	0.9914
DNN-At -F4	0.0023	0.0288	92.79	0.9910
DNN-At -F5	0.0022	0.0275	92.86	0.9916

Table 5.4: Numeric results of MSE, MAE, R^2, and AUC quality metrics assessing ML performance used with 2BDF-At features descriptor.

in the metric values, the performance throughout the five folds showed a similar pattern. A score of (0.0021, 0.0270, 93.51%, and 0.9946) for the metrics (MSE, MAE, R^2, and AUC, respectively) was achieved by Fold 1 demonstrating the highest performing model, followed closely by Fold 5.

5.5.2 Prediction results of 3BDF-At-based models

The three-body distribution function descriptor developed with the atomic approach produces a dataset defined as follows: every structure of the dataset is associated with a set of vectors (equal to the number of atoms in that structure), each containing 468 features.

This dataset went through the same 5-fold cross-validation resulting in 5 new training/testing sets. They were used as inputs to train and test the developed DNN-At model, and the testing phase results with regards to the quality metrics are reflected in Figure 5.9.

The inspection of the plots in Figure 5.9 affirms that the five folds of the 3BDF-At descriptor dataset slightly differ from one another in terms of energy prediction performance, which reflects a well-balanced data distribution. Moreover, the results illustrated in the graph representing the MSE, MAE and R^2 statistics in Figure 5.9 (a) are consistent with those presented in the ROC plot of Figure 5.9 (b).

The precise values of the assessment metrics with regards to the testing phase performance of the 3BDF-At's five folds are listed in Table 5.5.

As evidenced by Table 5.5, the performance of the 3BDF-At's five folds achieved results that are satisfactory for the energy prediction task. We notice a small variation between the different folds performance where Folds 3 and 4 yielded the best results. The best (MSE, MAE, R^2, and AUC) score of (0.0049, 0.0407, 83.07%, and 0.9742, respectively) was recorded by Fold 4.

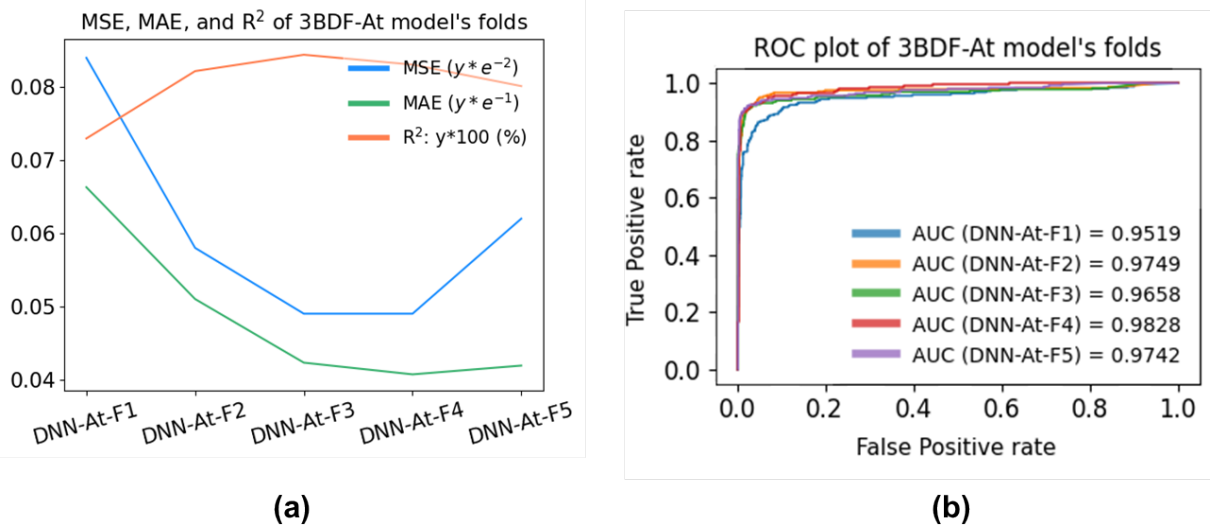


Figure 5.9: Accuracy assessment of energy prediction's testing phase of the 3BDF-At models' 5-folds. (a) MSE, MAE, R^2 measures, (b) ROC plot.

Model	MSE	MAE	R^2	AUC
DNN-At-F1	0.0084	0.0663	72.97	0.9519
DNN-At -F2	0.0058	0.0510	82.18	0.9749
DNN-At -F3	0.0049	0.0423	**84.42**	0.9658
DNN-At -F4	**0.0049**	**0.0407**	83.07	**0.9828**
DNN-At -F5	0.0062	0.0419	80.13	0.9742

Table 5.5: Numeric results of MSE, MAE, R^2, and AUC quality metrics assessing ML performance used with 3BDF-At features descriptor.

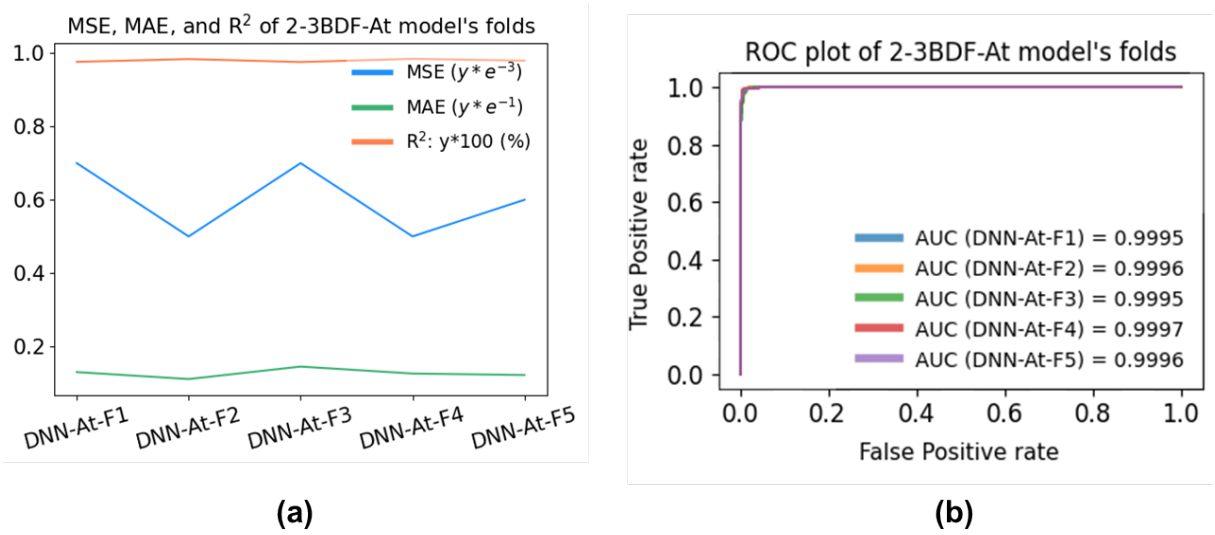


Figure 5.10: Accuracy assessment of energy prediction's testing phase of the 2-3BDF-At models' 5-folds. (a) MSE, MAE, R^2 measures, (b) ROC plot.

5.5.3 Prediction results of 2-3BDF-At-based models

The last dataset generated through the features engineering process in this study is the 2-3BDF-At. It uses the two- and three-body distribution functions with the atomic approach. It represents the largest dataset in terms of features with 528 size descriptor vectors for each atom in a structure.

After developing a DNN model according to this descriptor, the built model was trained and tested for the energy prediction. The five pairs of training/testing sets were generated with the 5-fold cross-validation process. Through Figure 5.10, a reflection of the 2-3BDF-At's folds performance is illustrated.

Figure 5.10 (a) represents a plot of the 2-3BDF-At's testing phase performance with regards to the statistical metrics MSE, MAE, and R^2. It shows that the five folds' results are very close to each other with tiny variations differentiating them. The ROC plot in Figure 5.10 (b) complies with the previous plot; it depicts a very narrow distinction between the 5 folds performances.

It is worth noting that the five folds of the 2-3BDF-At model yielded a solid performance and achieved roust results as reflected by the ROC curves demonstrating each a nearly perfect right-angled curve.

The quantitative values of the assessment metrics evaluating the performance of the 2-3BDF-At's five folds are presented in Table 5.6.

The 2-3BDF-At models, as shown in Table 5.6, attained commendable prediction ability

Model	MSE	MAE	R^2	AUC
DNN-At-F1	0.0007	0.0130	97.62	0.9995
DNN-At -F2	**0.0005**	**0.0111**	98.38	0.9996
DNN-At -F3	0.0007	0.0145	97.57	0.9995
DNN-At -F4	**0.0005**	0.0126	**98.44**	**0.9997**
DNN-At -F5	0.0006	0.0122	97.94	0.9996

Table 5.6: Numeric results of MSE, MAE, R^2, and AUC quality metrics assessing ML performance used with 2-3BDF-At features descriptor.

of the energy property. The evaluation of all folds indicates that similar results were obtained, with Fold 4 exhibiting a marginal advantage over the others, followed closely by Fold 3. The outperforming fold recorded a score of 0.0005, 0.0126, 98.44%, and 0.9997 for MSE, MAE, R^2, and AUC, respectively.

From the examination of this atomic approach, we perceive a familiar behavior of 2-3BDF-At's models outperforming the two other atomic-based models, followed by 2BDF-At's DNNs, whereas, the performance accuracy drops with the 3BDF-At's DNNs. Therefore, we conclude that 3BDF-At descriptor has a subpar data correlation compared to 2BDF-At and 2-3BDF-At descriptors.

5.6 Comparative analysis: Structural VS Atomic

The sole purpose of this study is to identify the best descriptor for crystal structures for the task of energy prediction. Selecting the outperforming descriptors matches the selection of the best ML model since the ML models in this study were developed correspondingly to support each descriptor.

Given the fact that there are six different descriptors (three in each approach), the best quality measures obtained from all descriptor models applied in terms of structural and atomic approaches were selected for the purpose of a comparative analysis.

Figure 5.11 illustrates the comparison between the best performing models selected, namely (2BDF-DNN-St-F4, 3BDF-DNN-St-F4, 2-3BDF-DNN-St-F1, 2BDF-At-F1, 3BDF-At-F4 and 2-3BDF-At-F4) applied to their corresponding structural and atomic descriptors. It reflects their predictive ability of the energy property, providing a clear overview of their relative strengths and weaknesses. The comparison is based on the same set of quality assessment parameters including (a) (MSE, MAE, and R^2) statistical measures which values

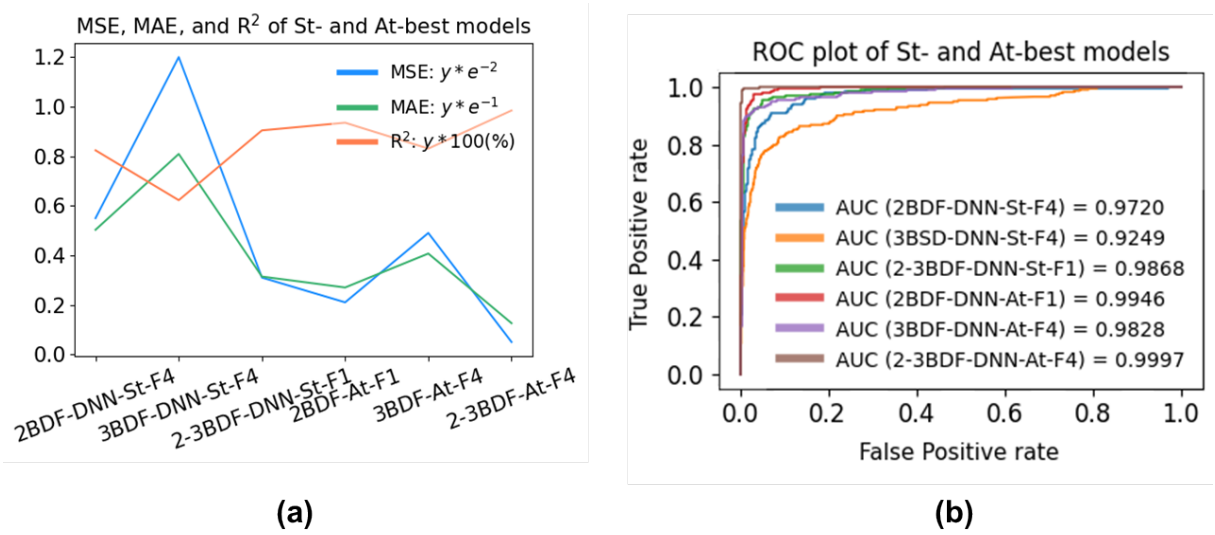


Figure 5.11: Performance comparison through: (a) MSE, MAE, R^2, and (b) ROC/AUC of the best selected models from both structural and atomic approaches.

can be obtained as explained in the legend of the plot, and (b) ROC/AUC graph.

The comparison plotted in Figure 5.11 proves that atomic-features-based models outperformed the structural-features-based ones. Also, in each approach, models implementing 2-3BDF descriptors yielded the best results, followed by 2BDF, then 3BDF as the least satisfactory models. In addition, the strongest and weakest performances were respectively presented by 2-3BDF's atomic-features-based model and 3BDF's structural-features-based model.

To summarize, the results show that, (1) regardless of the approach used, models implementing a combination of 2BDF and 3BDF are more efficient and better performing than models implementing 2BDF or 3BDF separately. (2) Moreover, 2-3BDF-At-F4 outperformed all other models in both approaches with the highest score on all measures. (3) The accuracy achieved of this model is due to the choice of the features descriptor. The combination of 2BDF and 3BDF is far more robust since it contains bonds and information about both pairs and triples of atoms. (4) Also, the atomic-based features extraction generates atom-wise descriptors, meaning that each structure has as many features descriptors as it has atoms; thus, these descriptors are considered very precise and accurate.

5.7 Validation

To validate 2-3BDF-At-F4 as superior to all other models, first, the model is applied to the entire dataset and then, on the unseen 20% of the remaining validation set in order to examine its behavior. Figure 5.12 illustrates the performance of the energy prediction when examining the selected model.

The subplot (e) of Figure 5.12 represents the 2-3BDF-At-F4 model's fitness of the full dataset. It is clearly noted that the selected model's output (in blue circles) matches the dataset's target (in red crosses) as they almost seamlessly overlap. In order to make sure that the dataset's target has the same pattern underneath the model's output, we plotted them separately in the subplots (a) and (c), respectively. This helped us further confirm that the model's output and the target set have nearly the same pattern.

The subplots (b) and (d) of Figure 5.12 respectively depict the statistical and graphical representations the 2-3BDF-At-F4 results with regards to the testing (as previously presented) and validation stages. We note that the model has a remarkable ability to generalize to new unseen data, as it yielded robust results in the validation stage that are not too distant from the testing phase results.

Finally Figure 5.12 (f) shows the regression plot of 2-3BDF-At-F4's validation stage of the energy prediction performance, where the target vs. predicted values of only the validation stage were illustrated in a plot demonstrating an arrangement of the points into a fitting line.

Results of the generalization and the validation stage proved the efficiency of the chosen features descriptor. The same set of measurements was used for the validation's evaluation; the obtained score was $1.1e^{-3}$, $1.69e^{-2}$, 96.44%, and 0.9976 for MSE, MAE, R^2, and AUC, respectively. Moreover, Table 5.7 presents the comparison of the target values ($y_i \in \mathbb{R}$) and output values ($\hat{y}_i \in \mathbb{R}$) generated by the best selected model for a randomly chosen sample of each database.

The performance of 2-3BDF-At-F4 shows through the similarity between the observed energy values and the predicted ones. Indeed, the difference between the normalized target and output energy values ranges from $9e^{-7}$ to $1e^{-5}$, while the average difference between the real target and output energy values is $8e^{-4}$.

5.8 Comparison with the state of the art

Conducting a comparative analysis of a study with the state of the art can be crucial since it offers a foundation for validating the quality and the effectiveness of the proposed approach.

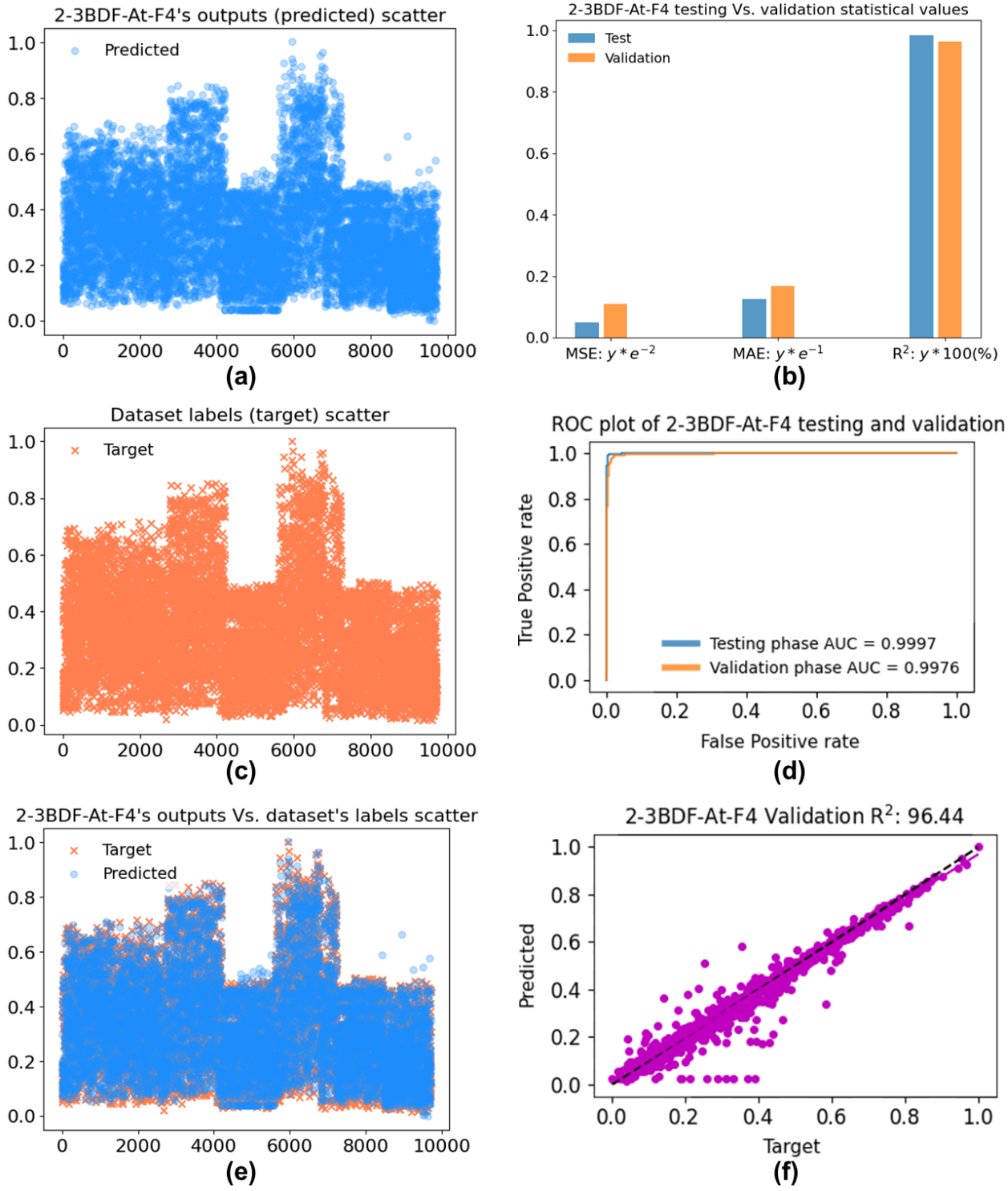


Figure 5.12: Performance of the best selected model 2-3BDF-At-F4 for the validation stage. (a, c, e) predicted Vs. targets of the entire dataset, (b, d) comparison between testing and validation, (f) regression plot of the model on the validation dataset.

Database sample	Normalized target	Normalized output	Target	Output
$Li_{20}Ag_2$	0.3589	0.3589	−40.0670	−40.0673
Li_8At_{18}	0.4871	0.4871	−54.3876	−54.3859
$Li_{16}Au_{14}$	0.7994	0.7994	−89.2555	−89.2563
Li_6Ba	0.1038	0.1038	−11.5891	−11.5910
$Li_{10}Bi_2$	0.2449	0.2449	−27.3397	−27.3398
Li_6Br_{10}	0.4106	0.4106	−45.8399	−45.8414
$Li_{18}Ca_4$	0.3578	0.3578	−39.9458	−39.9457
Li_9Cd_5	0.1795	0.1795	−20.0450	−20.0454

Table 5.7: Comparing a sample of each database's target value with the according output value generated by 2-3BDF-At-F4.

However, without a proper scale of comparison, the analysis of the state of the art vs. the proposed approach of a study would be unfair. Most of the studies of crystal structure prediction present unique investigations with respect to datasets, prediction task (classification or regression), desired property to be predicted, and performance metrics. In our study, we investigated the prediction of the energy property (regression task) of the databases LiAg, LiAt, LiAu, LiBa, LiBi, LiBr, LiCa, and LiCd. The validation of the presented approach was performed locally by applying the built selected model on 20% of the total merged datasets that was never introduced in the training or testing phases of the model. The performance of the model was measured using MSE, MAE, R^2, and ROC/AUC metrics. To our best knowledge, a study of these exact databases for the energy property prediction has not been published before, thus, making our study a unique investigation.

Nevertheless, in order to have a better idea of where this study stands in terms of effectiveness of its predictive ability and superiority of its innovation, it is still possible to compare its results with some of state-of-the-art works. The selected relevant studies to compare with are the ones which focused on the energy property prediction and where the evaluation process was conveyed using at least one of the quality metrics used in this study (see Table 5.8).

Table 5.8 demonstrates that the approach proposed in this study for the task of energy prediction achieved better results than the selected state-of-the-art studies.

Ref.	Data type	Data source	ML model	Descriptor	MAE
[55] 2015	Comput.	MP	KRR	Ewald sum-based CM, Extended CM, Sine matrix	0.07
[87] 2018	Experim.	ICSD	GNN-based CGCNN	Defined	0.039
[88] 2020	Comput.	MP	GNN-based GATGNN	Defined	0.039
[89] 2020	Comput.	Generated	GNN-based OGCNN	OFM	0.0466
[90] 2022	Comput., experim.	OQDM, MatB	GNN-based	Defined	0.016
Our work	Comput.	Generated	DNN-At	Defined (2-3BDF-At)	0.0126

Table 5.8: Comparison of 2-3BDF-At performance with that of the state of the art in terms of energy prediction and MSE measure.

5.9 Conclusion

In conclusion, we have revealed, in this chapter, the findings of our quest to predict the energy property of crystal structures which holds the key to understanding the stability of these materials and their behavior. Crystallography, features engineering, and machine / deep learning (explored in the previous chapters) all cooperated to predict crystal structural energy with remarkable accuracy, illuminating the behavior of crystals. Beyond these numbers, there are significant implications and insights provided in a way to guide the future investigations of materials research and discovery.

This chapter first explored the strategy and the environment in which the selected machine learning models were implemented. Then, the results of each model with every descriptor of the two approaches (structural and atomic) were presented. We displayed the results in a way to compare approaches, descriptors, and models in order to 1) approve of the strongest descriptor (making the approach to which it belongs to the strongest of the two), and 2) validate the best performing model for crystal structure energy prediction. It has been determined by the previously demonstrated results in this chapter that the 2-3BDF-At-F4 DNN model outperformed all other models, indicating that the combination of the two- and three-body distribution functions with the atomic approach is considered to be the superior

descriptor, and that the unconventional deep neural network model developed is the best among the implemented ML/DL models.

General Conclusion

In the present thesis, we provided an in-depth investigation of crystal structure energy prediction through the lens of machine learning, covering a wide range of computational methods and materials science. Our study has been driven by the core objective of advancing materials science through the accurate regression of the energy property within crystalline structures. We have dived into the complexities of crystallography, revealing the fundamental relationships that govern energy behaviors through diligent and meticulous features engineering, algorithmic innovation, and rigorous model validation. This research marks a ground-breaking effort that advances our understanding of crystalline matter. Moreover, the insights gained from our study empower the search for new materials with tailored properties and redefine the future of materials discovery, eventually advancing materials science and engineering.

In order to meet our main goals and objectives, the path to this journey was marked by several key phases. As a start, we methodically outlined the scope of our research, pinpointing the critical intersection of crystallography, machine learning, and energy property in materials. During this stage, a thorough examination of the existing literature was conducted to help us shape our research and formulate our study questions and goals.

With a clear direction, we began with the data collection process, using reliable crystallographic databases. In order to assure data consistency and quality, this step entailed meticulous data preparation, laying the groundwork for further analysis. The main data preprocessing and preparation process was performed using simple NLP techniques.

Having preprocessed data at hand, we proceeded with converting complex structural information into machine-readable inputs for our models. In this critical phase, we proposed two approaches to describe crystal structures. The former structural approach defines through a numerical vector a whole crystal structure and is therefore convenient for any learning process. The latter atomic approach however, offers an atom-wise representation, thus defining a crystal structure with as many vectors as it has atoms making it non-uniform and unfit for the (direct) learning process. This pivotal phase addressed a key objective of our study, that is crystal structure features engineering. The resulting features were further analyzed, and

it was unveiled that the relationship between input descriptors and the energy property is of a non-linear nature. The acquired information of this analysis is extremely important as an insightful perception of the machine learning modeling to be performed in the next step.

The heart of our research lay in the development and validation of machine learning models. We trained and tested a wide range of algorithms through two approaches in accordance with the features engineering process. A total of five different models were used for the structural modeling approach, including four machine learning-based algorithms, namely: ElasticNet, Bayesian ridge, random forest, and support vector machine, in addition to a deep learning-based artificial neural network. In the second, atomic modeling approach, since the corresponding atom-wise representation is non-uniform, we proposed and developed a non-conventional deep learning-based artificial neural network topology to support these descriptors. The models from both approaches were refined through hyper-parameter tuning to optimize their prediction skills.

A thorough evaluation process was performed to assess the accuracy of the investigated ML algorithms using 5-fold cross-validation technique and performance measures, including MSE, MAE, R^2, and ROC/AUC. Through a sequence of graphical representations and numerical data, we illustrated and analyzed the achieved results. It has been determined from the examination of our findings that the proposed atomic modeling approach's deep neural network outperformed all other models with regards to every performance metrics for the energy property prediction. In terms of descriptors, the most robust crystallographic representation was identified to be the combination of atomic two- and three-body distributions functions. Moreover, compared to recent studies of the state of the art, our work yielded remarkable crystal structure energy prediction results.

Throughout this research process, we faced and overcame challenges, adopted novel approaches, and, most importantly, improved our comprehension of crystal structure energy prediction. The closing of this thesis marks the conclusion of a research endeavor on one hand, but the beginning of a future shaped by insights uncovered in the field of crystal structure prediction. Our future works intend to use the best identified 2-3BDF-At descriptor with its developed unconventional DNN model on larger sets of data for other properties and thus, broader applications. Moreover, it is worth mentioning that the computational complexity of deep learning-based models is particularly high and increases with the number of data and features. Consequently, in order to avoid the necessity of having huge amount of data for the modeling process, reducing the computational cost, we seek to develop a more efficient learning-free regression approach for crystal structure property prediction.

Bibliography

[1] Sarah L Price. Predicting crystal structures of organic compounds. *Chemical Society Reviews*, 43(7):2098–2111, 2014.

[2] Hugh PG Thompson and Graeme M Day. Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science*, 5(8):3173–3182, 2014.

[3] Kun Wang, DY Li, Zejie Fei, Xianfeng Ma, and Xiaoqin Zeng. Discovery of a new crystal structure of $LiBeF_3$ and its thermodynamic and optical properties. *Computational Materials Science*, 169:109077, 2019.

[4] Marc De Graef and Michael E McHenry. *Structure of materials: an introduction to crystallography, diffraction and symmetry*. Cambridge University Press, 2012.

[5] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

[6] Stefano Curtarolo, Dane Morgan, Kristin Persson, John Rodgers, and Gerbrand Ceder. Predicting crystal structures with data mining of quantum calculations. *Physical review letters*, 91(13):135503, 2003.

[7] Alex Smola. Introduction to machine learning, 2008.

[8] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):83, 2019.

[9] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, and Kristin Persson. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. *Journal of Materials Research*, 31(8):977–994, 2016.

[10] Richard A Dunlap. *Novel Microstructures for Solids*. Morgan & Claypool Publishers Bristol, 2018.

[11] Aloke Paul, Tomi Laurila, Vesa Vuorinen, and Sergiy V Divinski. *Thermodynamics, diffusion and the Kirkendall effect in solids*. Springer, 2014.

[12] Guy Crundwell. Crystal structures: Lattices and solids in stereoview (by mark ladd). *Journal of Chemical Education*, 77(12):1563, 2000.

[13] Richard JD Tilley. *Crystals and crystal structures*. John Wiley & Sons, 2020.

[14] Walter Borchardt-Ott. *Crystallography: an introduction*. Springer Science & Business Media, 2011.

[15] Mariusz Jaskolski, Zbigniew Dauter, and Alexander Wlodawer. A brief history of macromolecular crystallography, illustrated by a family tree and its nobel fruits. *The FEBS journal*, 281(18):3985–4009, 2014.

[16] Michael Eckert. Max von laue and the discovery of x-ray diffraction in 1912, 2012.

[17] John Meurig Thomas. The birth of x-ray crystallography. *Nature*, 491(7423):186–187, 2012.

[18] John P Perdew and Adrienn Ruzsinszky. Fourteen easy lessons in density functional theory. *International Journal of Quantum Chemistry*, 110(15):2801–2807, 2010.

[19] Nick Laskin. Fractional schrödinger equation. *Physical Review E*, 66(5):056108, 2002.

[20] Kieron Burke. Perspective on density functional theory. *The Journal of chemical physics*, 136(15):150901, 2012.

[21] Clemence Corminboeuf, Fabien Tran, and Jacques Weber. The role of density functional theory in chemistry: Some historical landmarks and applications to zeolites. *Journal of Molecular Structure: THEOCHEM*, 762(1-3):1–7, 2006.

[22] Robert O Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of modern physics*, 87(3):897, 2015.

[23] Axel D Becke. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of chemical physics*, 140(18):18A301, 2014.

[24] Di Zhou. An introduction of density functional theory and its application. *Physics. Drexel. Edu*, 2007.

[25] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017.

[26] AP Othman, Amin Aadenan, Muhammad Mus Ab Anas, and GA Gopir. Calculation of electron density distribution in $Cd_{0.5}Zn_{0.5}S$ by density functional theory: Exploring its bonding character and stability. In *Advanced Materials Research*, volume 1108, pages 21–25. Trans Tech Publ, 2015.

[27] Philip J Hasnip, Keith Refson, Matt IJ Probert, Jonathan R Yates, Stewart J Clark, and Chris J Pickard. Density functional theory in the solid state. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011):20130270, 2014.

[28] Mohamed Barhoumi. The density functional theory and beyond: Example and applications. In *Density Functional Theory-Recent Advances, New Perspectives and Applications*. IntechOpen, 2021.

[29] A Van de Walle and G Ceder. Correcting overbinding in local-density-approximation calculations. *Physical Review B*, 59(23):14992, 1999.

[30] Xinlei Hua, Xiaojie Chen, and WA Goddard. Generalized generalized gradient approximation: An improved density-functional theory for accurate orbital eigenvalues. *Physical Review B*, 55(24):16103, 1997.

[31] Yi X Wang, Hua Y Geng, Q Wu, and Xiang R Chen. Orbital localization error of density functional theory in shear properties of vanadium and niobium. *The Journal of Chemical Physics*, 152(2):024118, 2020.

[32] Abdelkrim Mostefai. Charge density and density of states (DOS) of monoclinic ZrO_2 using Meta-GGA DFT functional. 2022.

[33] Dayou Zhang and Donald G Truhlar. Unmasking static correlation error in hybrid kohn–sham density functional theory. *Journal of Chemical Theory and Computation*, 16(9):5432–5440, 2020.

[34] Jack D Dunitz. Are crystal structures predictable? *Chemical Communications*, (5):545–548, 2003.

[35] Sławomir J Grabowski. *Hydrogen bonding: new insights*, volume 3. Springer, 2006.

[36] Eray S Aydil. Nanomaterials for solar cells. *Nanotech. L. & Bus.*, 4:275, 2007.

[37] Hania Djani. *Modélisation ab initio des oxydes ferroélectriques à phase aurivillius*. PhD thesis, 2013.

[38] VA Isupov. Anomalous properties of $A_{m-1}Bi_2B_mO_{3m+3}$ layered ferroelectrics. *Inorganic materials*, 42(11):1236–1242, 2006.

[39] Kurt R Kendall, Carlos Navas, Julie K Thomas, and Hans-Conrad zur Loye. Recent developments in oxide ion conductors: Aurivillius phases. *Chemistry of materials*, 8(3):642–649, 1996.

[40] VA Isupov. Systematization of aurivillius-type layered oxides. *Inorganic materials*, 42(10):1094–1098, 2006.

[41] Yanhui Wang, Emilie Delahaye, Cedric Leuvrey, Fabrice Leroux, Pierre Rabu, and Guillaume Rogez. Post-synthesis modification of the aurivillius phase $Bi_2SrTa_2O_9$ via in situ microwave-assisted "click reaction". *Inorganic Chemistry*, 55(19):9790–9797, 2016.

[42] MS Tomar. Structural and ferroelectric properties of aurivillius phase materials. *Integrated Ferroelectrics*, 42(1):191–205, 2002.

[43] Claudia Milena Bedoya Hincapie, Manuel Jonathan Pinzon Cardenas, Jose Edgar Alfonso Orjuela, Elisabeth Restrepo Parra, and Jhon Jairo Olaya Florez. Physical-chemical properties of bismuth and bismuth oxides: Synthesis, characterization and applications. *Dyna*, 79(176):139–148, 2012.

[44] Bruno A Calfa and John R Kitchin. Property prediction of crystalline solids from composition and crystal structure. *AIChE Journal*, 62(8):2605–2613, 2016.

[45] Woon Bae Park, Jiyong Chung, Jaeyoung Jung, Keemin Sohn, Satendra Pal Singh, Myoungho Pyo, Namsoo Shin, and K-S Sohn. Classification of crystal structure using a convolutional neural network. *IUCrJ*, 4(4):486–494, 2017.

[46] Jake Graser, Steven K Kauwe, and Taylor D Sparks. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chemistry of Materials*, 30(11):3601–3612, 2018.

[47] Gerbrand Ceder, Dane Morgan, Chris Fischer, Kevin Tibbetts, and Stefano Curtarolo. Data-mining-driven quantum mechanics for the prediction of structure. *MRS bulletin*, 31(12):981–985, 2006.

[48] Evgeny V Podryabinkin, Evgeny V Tikhonov, Alexander V Shapeev, and Artem R Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6):064114, 2019.

[49] Artem R Oganov. Crystal structure prediction: reflections on present status and challenges. *Faraday discussions*, 211:643–660, 2018.

[50] Kristof T Schütt, Henning Glawe, Felix Brockherde, Antonio Sanna, Klaus-Robert Müller, and Eberhard KU Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B*, 89(20):205118, 2014.

[51] Meriem Mouzai, Saliha Oukid, and Aouache Mustapha. Machine learning modeling for the prediction of materials energy. *Neural Computing and Applications*, pages 1–18, 2022.

[52] Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18:241–252, 2020.

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

[54] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O Anatole Von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Muller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

[55] Felix Faber, Alexander Lindmaa, O Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, 2015.

[56] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[57] G Bergerhoff, ID Brown, F Allen, et al. Crystallographic databases. *International Union of Crystallography, Chester*, 360:77–95, 1987.

[58] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017.

[59] Aleksandr V Fedorov and Ivan V Shamanaev. Crystal structure representation for neural networks using topological approach. *Molecular Informatics*, 36(8):1600162, 2017.

[60] Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quiros, Nadezhda R Serebryanaya, Peter Moeck, Robert T Downs, and Armel Le Bail. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic acids research*, 40(D1):D420–D427, 2012.

[61] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):1–12, 2017.

[62] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H Taylor, Lance J Nelson, Gus LW Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, et al. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012.

[63] Mahbub Hussain, Jordan J Bird, and Diego R Faria. A study on cnn transfer learning for image classification. In *UK Workshop on computational Intelligence*, pages 191–202. Springer, 2018.

[64] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli. Insightful classification of crystal structures using deep learning. *Nature communications*, 9(1):1–10, 2018.

[65] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jorg Behler, Gábor Csányi, Alexander V Shapeev, Aidan P Thompson, Mitchell A Wood, et al. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A*, 124(4):731–745, 2020.

[66] Y Mishin. Machine-learning interatomic potentials for materials science. *Acta Materialia*, 214:116980, 2021.

[67] Sergey Pozdnyakov, Artem R. Oganov, Efim Mazhnik, Arslan Mazitov, and Ivan Kruglov. Fast general two- and three-body interatomic potential. *Phys. Rev. B*, 107:125160, Mar 2023.

[68] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, 140(32):10158–10168, 2018.

[69] Wen Tong, Qun Wei, Hai-Yan Yan, Mei-Guang Zhang, and Xuan-Min Zhu. Accelerating inverse crystal structure prediction by machine learning: A case study of carbon allotropes. *Frontiers of Physics*, 15(6):1–7, 2020.

[70] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications*, 183(10):2063–2070, 2012.

[71] Andrey A Golov, Artem A Kabanov, Davide M Proserpio, and Vladislav A Blatov. SACADA-the database of three periodic carbon allotropes. In *ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES*, volume 71, pages S356–S356. INT UNION CRYSTALLOGRAPHY 2 ABBEY SQ, CHESTER, CH1 2HU, ENGLAND, 2015.

[72] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[73] Maximilian Amsler, Logan Ward, Vinay I Hegde, Maarten G Goesten, Xia Yi, and Chris Wolverton. Ternary mixed-anion semiconductors with tunable band gaps

from machine-learning and crystal structure prediction. *Physical Review Materials*, 3(3):035404, 2019.

[74] Keisuke Takahashi and Lauren Takahashi. Creating machine learning-driven material recipes based on crystal structure. *The journal of physical chemistry letters*, 10(2):283–288, 2019.

[75] Santosh Behara, Taher Poonawala, and Tiju Thomas. Crystal structure classification in ABO_3 perovskites via machine learning. *Computational Materials Science*, 188:110191, 2021.

[76] Yuxin Li, Rongzhi Dong, Wenhui Yang, and Jianjun Hu. Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors. *Computational Materials Science*, 198:110686, 2021.

[77] Yong Zhao, Yuxin Cui, Zheng Xiong, Jing Jin, Zhonghao Liu, Rongzhi Dong, and Jianjun Hu. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. *ACS omega*, 5(7):3596–3606, 2020.

[78] Yuta Suzuki, Hideitsu Hino, Yasuo Takeichi, Takafumi Hawai, Masato Kotsugi, and Kanta Ono. Machine learning-based crystal structure prediction for x-ray microdiffraction. *Microscopy and Microanalysis*, 24(S2):142–143, 2018.

[79] Yuta Suzuki, Hideitsu Hino, Takafumi Hawai, Kotaro Saito, Masato Kotsugi, and Kanta Ono. Symmetry prediction and knowledge discovery from x-ray diffraction patterns using an interpretable machine learning approach. *Scientific reports*, 10(1):1–11, 2020.

[80] Taewon Jin, Ina Park, Taesu Park, Jaesik Park, and Ji Hoon Shim. Accelerated crystal structure prediction of multi-elements random alloy using expandable features. *Scientific reports*, 11(1):1–9, 2021.

[81] Sams Jarin, Yufan Yuan, Mingxing Zhang, Mingwei Hu, Masud Rana, Sen Wang, and Ruth Knibbe. Predicting the crystal structure and lattice parameters of the perovskite materials via different machine learning models based on basic atom properties. *Crystals*, 12(11):1570, 2022.

[82] Jessica M. Hudspeth. Short-range order in ferroelectric triglycine sulphate. *Acta Crystallographica Section A*, 67:416–416, 2011.

[83] Minoru Kusaba, Chang Liu, and Ryo Yoshida. Crystal structure prediction with machine learning-based element substitution. *Computational Materials Science*, 211:111496, 2022.

[84] Abhik Chakraborty and Raksha Sharma. A deep crystal structure identification system for x-ray diffraction patterns. *The Visual Computer*, 38(4):1275–1282, 2022.

[85] Jianjun Hu, Wenhui Yang, Rongzhi Dong, Yuxin Li, Xiang Li, Shaobo Li, and Edirisuriya M. D. Siriwardane. Contact map based crystal structure prediction using global optimization. *CrystEngComm*, 23:1765–1776, 2021.

[86] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[87] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[88] Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148, 2020.

[89] Mohammadreza Karamad, Rishikesh Magar, Yuting Shi, Samira Siahrostami, Ian D Gates, and Amir Barati Farimani. Orbital graph convolutional neural network for material property prediction. *Physical Review Materials*, 4(9):093801, 2020.

[90] Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. Crystal structure prediction by combining graph network and optimization algorithm. *Nature communications*, 13(1):1–8, 2022.

[91] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):1–10, 2020.

[92] Haidi Wang, Yuzhi Zhang, Linfeng Zhang, and Han Wang. Crystal structure prediction of binary alloys via deep potential. *Frontiers in chemistry*, 8:589795, 2020.

[93] Changho Hong, Jeong Min Choi, Wonseok Jeong, Sungwoo Kang, Suyeon Ju, Kyeongpung Lee, Jisu Jung, Yong Youn, and Seungwu Han. Training machine-learning potentials for crystal structure prediction using disordered structures. *Physical Review B*, 102(22):224104, 2020.

[94] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[95] Alexander V Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.

[96] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[97] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. USPEX—Evolutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.

[98] Andriy O Lyakhov, Artem R Oganov, Harold T Stokes, and Qiang Zhu. New developments in evolutionary structure prediction algorithm USPEX. *Computer Physics Communications*, 184(4):1172–1182, 2013.

[99] Artem R Oganov, Yanming Ma, Colin W Glass, and Mario Valle. Evolutionary crystal structure prediction: overview of the USPEX method and some of its applications. *Psi-k Newsletter*, 84:142–171, 2007.

[100] Lauri Himanen, Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[101] Atsuto Seko, Atsushi Togo, and Isao Tanaka. Descriptors for machine learning of materials data. In *Nanoinformatics*, pages 3–23. Springer, Singapore, 2018.

[102] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013.

[103] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[104] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[105] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[106] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[107] Simon Parsons. Introduction to machine learning, second editon by ethem alpaydin, mit press, 584 pp., $55.00. isbn 978-0-262-01243-0. *The Knowledge Engineering Review*, 25(3):353–353, 2010.

[108] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[109] Inga M Müller. Feature selection for energy system modeling: Identification of relevant time series information. *Energy and AI*, 4:100057, 2021.

[110] Hamid Gholami, Aliakbar Mohamadifar, Armin Sorooshian, and John D Jansen. Machine-learning algorithms for predicting land susceptibility to dust emissions: The case of the Jazmurian Basin, Iran. *Atmospheric Pollution Research*, 11(8):1303–1315, 2020.

[111] A George Assaf, Mike Tsionas, and Anastasios Tasiopoulos. Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tourism Management*, 71:1–8, 2019.

[112] Achmad Efendi and Effrihan. A simulation study on Bayesian Ridge regression models for several collinearity levels. In *AIP conference proceedings*, volume 1913, page 020031. AIP Publishing LLC, 2017.

[113] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[114] Jihed Khiari, Luis Moreira-Matias, Ammar Shaker, Bernard Ženko, and Sašo Džeroski. Metabags: Bagged meta-decision trees for regression. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 637–652. Springer, 2019.

[115] Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013.

[116] Helmut Schmid. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154, 2013.

[117] Johannes L Grabmeier and Larry A Lambe. Decision trees for binary classification variables grow equally with the Gini impurity measure and pearson's chi-square test. *International journal of business intelligence and data mining*, 2(2):213–226, 2007.

[118] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[119] Simon Bernard, Laurent Heutte, and Sebastien Adam. On the selection of decision trees in random forests. In *2009 International Joint Conference on Neural Networks*, pages 302–307. IEEE, 2009.

[120] Balraj Singh, Parveen Sihag, and Karan Singh. Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Modeling Earth Systems and Environment*, 3:999–1004, 2017.

[121] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[122] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[123] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.

[124] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, pages 758–763. Springer, 2019.

[125] Mathieu Wauters and Mario Vanhoucke. Support vector machine regression for project control forecasting. *Automation in Construction*, 47:92–106, 2014.

[126] Li Zhang, Weida Zhou, and Licheng Jiao. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):34–39, 2004.

[127] Theodore B Trafalis and Huseyin Ince. Support vector machine for regression and applications to financial forecasting. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 6, pages 348–353. IEEE, 2000.

[128] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[129] Enzo Grossi and Massimo Buscema. Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, 19(12):1046–1054, 2007.

[130] Raheel Zafar, Sarat C Dass, and Aamir Saeed Malik. Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion. *PloS one*, 12(5):e0178410, 2017.

[131] Andrei Dobrescu, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Understanding deep neural networks for regression in leaf counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[132] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2013.

[133] Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michele Sebag. Collaborative hyperparameter tuning. In *International conference on machine learning*, pages 199–207. PMLR, 2013.

[134] Mohamad Aqib Haqmi Abas, Nurlaila Ismail, NA Ali, S Tajuddin, and Nooritawati Md Tahir. Agarwood oil quality classification using support vector classifier and grid search cross validation hyperparameter tuning. *Int. J*, 8, 2020.

[135] Yongbin Yu, Kwabena Adu, Nyima Tashi, Patrick Anokye, Xiangxiang Wang, and Mighty Abra Ayidzoe. RMAF: Relu-Memristor-Like Activation Function for deep learning. *IEEE Access*, 8:72727–72741, 2020.

[136] H Altun, A Bilgil, and BC Fidan. Treatment of multi-dimensional data to enhance neural network estimators in regression problems. *Expert Systems with Applications*, 32(2):599–605, 2007.

[137] Jin Huang and Charles X Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[138] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2020.

[139] Meftah Salem M Alfatni, Abdul Rashid Mohamed Shariff, Siti Khairunniza Bejo, Osama M Ben Saaed, and Aouache Mustapha. Real-time oil palm FFB ripeness grading system based on ann, knn and svm classifiers. In *IOP conference series: earth and environmental science*, volume 169, page 012067. IOP Publishing, 2018.

[140] Aouache Mustapha, Aini Hussain, and Salina Abdul Samad. A new approach for noise reduction in spine radiograph images using a non-linear contrast adjustment scheme based adaptive factor. *Sci. Res. Essays*, 6(20):4246–4258, 2011.

[141] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[142] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[143] Yi Liu, Laijun Sun, Chengsi Du, and Xing Wang. Near-infrared prediction of edible oil frying times based on Bayesian Ridge Regression. *Optik*, 218:164950, 2020.

[144] Eiiti Kasuya. On the use of r and r squared in correlation and regression. Technical report, Wiley Online Library, 2019.

[145] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *Osdi*, volume 16, pages 265–283. Savannah, GA, USA, 2016.

[146] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[147] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[148] John D Hunter. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[149] B Prijamboedi, AA Nugroho, N Mufti, A Fajar, TTM Palstra, et al. Aurivillius phases of $PbBi_4Ti_4O_{15}$ doped with Mn^{3+} synthesized by molten salt technique: structure, dielectric, and magnetic properties. *Journal of solid state chemistry*, 184(5):1318–1323, 2011.

[150] A Peláiz-Barranco and Y González-Abreu. Ferroelectric ceramic materials of the aurivillius family. *Journal of Advanced Dielectrics*, 3(04):1330003, 2013.

[151] Eric J Nichols. *Aurivillius Phase Oxides for Photocatalytic Applications*. PhD thesis, Alfred University, 2010.

[152] EA Fortalnova, MG Safronenko, IA Smagin, ED Politova, MN Kurasova, and AV Mosunov. Synthesis and investigation of RE (III) cation substituted SBN and SBT ceramics. *Ferroelectrics*, 511(1):62–68, 2017.

[153] Justin Bergmann, Esko Oksanen, and Ulf Ryde. Combining crystallography with quantum mechanics. *Current Opinion in Structural Biology*, 72:18–26, 2022.

[154] Alessandro Genoni, Lukas Bučinskỳ, Nicolas Claiser, Julia Contreras-García, Birger Dittrich, Paulina M Dominiak, Enrique Espinosa, Carlo Gatti, Paolo Giannozzi, Jean-Michel Gillet, et al. Quantum crystallography: Current developments and future perspectives. *Chemistry–A European Journal*, 24(43):10881–10905, 2018.

[155] Simon Grabowsky, Alessandro Genoni, and Hans-Beat Bürgi. Quantum crystallography. *Chemical Science*, 8(6):4159–4176, 2017.

[156] B Winkler. An introduction to "computational crystallography". *Zeitschrift für Kristallographie-Crystalline Materials*, 214(9):506–527, 1999.

[157] Feliks Aleksandrovich Berezin and Mikhail Shubin. *The Schrödinger Equation*, volume 66. Springer Science & Business Media, 2012.

[158] William D Callister. An introduction: material science and engineering. *New York*, 106:139, 2007.

[159] Artem R Oganov. *Modern methods of crystal structure prediction*. John Wiley & Sons, 2011.

[160] Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of chemical information and modeling*, 60(10):4518–4535, 2020.