Predicting Materials Behavior with Atomistic Simulations and Machine Learning

Ayana Ghosh

*University of Connecticut - Storrs, ayana.ghosh@uconn.edu*

Follow this and additional works at: [https://opencommons.uconn.edu/dissertations](https://opencommons.uconn.edu/dissertations)

**Recommended Citation**


[https://opencommons.uconn.edu/dissertations/2604](https://opencommons.uconn.edu/dissertations/2604)
Data driven approaches based on machine learning (ML) algorithms are very popular in the domain of physical sciences for the determination of yet unknown structure-properties-performance relationships for a wide range of different material families. This dissertation focuses on studying a number of such cases where various ML algorithms and statistical techniques, coupled with appropriate materials data obtained from experiments and atomistic simulations, are employed to build comprehensive ML-based frameworks capable of predicting complex materials behavior. The materials spaces investigated encompass isolated organic molecules, polymer crystals, inorganic multiferroics and actinides, while the target system characteristics or functionalities include molecular crystallization propensity, ferroelectricity and magnetism, which are in turn connected to the structural and electronic properties of the considered materials. In order to gain electronic-level understanding (Human Learning) of functionalities, such as ferroelectricity and magnetism, we have examined four different systems using density functional theory (DFT) approaches. These studies provided sufficient introductory knowledge for construction of targeted, data-driven ML-based frameworks — described in this dissertation — for further evaluation of the materials properties of interest, as well as for prediction of novel materials with similar or advanced characteristics.
Predicting Materials Behavior with Atomistic Simulations and Machine Learning

Ayana Ghosh

B.S., University of Michigan, 2015
M.S., University of Connecticut, 2019

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2020
Copyright by
Ayana Ghosh

2020
Dedicated to my parents:

I am forever indebted to them for their unwavering support and blessings for me to pursue a career in basic sciences.
Acknowledgments

I would like to acknowledge every studious contribution of Prof. Serge Nakhmanson as my Major Advisor throughout my graduate career.

I am grateful to Prof. S. Pamir Alpay, Prof. Alexandru Asandei, Prof. Pu-Xian Gao, Dr. Geoffrey Wood and Dr. Jian-Xin Zhu for their help and guidance in the process of planning and writing this dissertation.

My special appreciations go to my primary mentor Dr. Jian-Xin Zhu for his excellent guidance, and to Dr. Shizeng Lin, Dr. Filip Ronning, Dr. Nicholas Lubbers and Dr. Hongchul Choi for their scientific insight and technical help during my concurrent visits to Los Alamos National Laboratory (LANL).

I am thankful to all my co-authors and collaborators: Dr. Lydie Louis and Dr. Krishna Pitike (UConn), Dr. Kapildev Arora, Dr. Bruno Hancock, Dr. Joseph Krzyzaniak and Dr. Paul Meenan (Pfizer), Dr. Towfiq Ahmed and Dr. Dzmitry A Yarotski (LANL), as well as Dennis Trujillo and Elizabeth Schiesser (UConn), for their multiple contributions to all the projects that are now included in this dissertation.

Funding sources: Pfizer Inc., Los Alamos National Laboratory, Air Force Research Laboratory.

Personal Note: I am fortunate to have friends and colleagues like Dennis Trujillo, Catherine Barachkova, Lorri Lafontaine, Mohamad Daeipour, Uche Anene, Celia Einhorn, Drew Einhorn, Arshia Mirza, Akash Vardhan Gundlapalli, Benzamin Frye, Yesha
Bhatt, Preshit Tambey, Tobi Adibesi, Rajarshi Nandi, Auditi Pramanick, Latasri Hazarika, Gautam Basu, Baby Basu, Soma Mandal, Kuntal Mandal, Philip Sanchez, Eliza Sanchez, Melisha Perera, Malin Clark, Dr. Syamaprasad Bhattacharya, Dr. Suchita Nandy Mazumder, Dr. Nalin Fernando, Dr. Shanta Thoutam, Dr. Patrice Cambell, Dr. Alan Grafe, Dr. Rajib Ganugly, Dr. Michelle Stark, Dr. James Alsup, Dr. Lukasz Kuna, Dr. Debayon Paul, Dr. Vignesh Vasu, Dr. Christopher Lane, Dr. Roxanne Tutchton, Dr. Jacqueline Garafano, Dr. Stefan Zollner, Dr. Subhashis Hazarika, Dr. Navamita Ray, Dr. Srirupa Chakraborty, Dr. Soumya Dutta, Dr. Shailja Jakhar, Dr. Partha Ghosal, Dr. Arup Ranjan Bhattacharya, Dr. Anirban Basu, Dr. Aparajita Banerjee and many others, located across the globe who have encouraged and supported me in my academic endeavors.

Last, but not the least, I would like to thank my parents, Ila Ghosh and Asim Kumar Ghosh, who have always inspired me to do better and not to give up under difficult circumstances. Only very special parents can let their daughter fly to a different continent at the age of 18 to let her fulfill her own wishes, without having much knowledge of what to expect on the other side of the globe. I am equally grateful and proud to have such souls in my life whose lifelong sacrifices have built myself a platform from where I began my journey. They are strong believers of hard work, honesty, integrity, persistence and these are the virtues I carry with myself to pursue a career of my dreams.

Couple of quotes out of many that have kept me motivated in my academic pursuit:

"Anything is possible, no matter how big the goal or how big the dream." – Michael Phelps

"Luck is indeed where preparation meets opportunity." – Dr. Randy Pausch
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Computational Methods</td>
<td>15</td>
</tr>
<tr>
<td>2.1</td>
<td>Density Functional Theory</td>
<td>16</td>
</tr>
<tr>
<td>2.1.1</td>
<td>General Considerations</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Kohn-Sham Equations</td>
<td>18</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Further Approximations</td>
<td>22</td>
</tr>
<tr>
<td>2.1.4</td>
<td>System periodicity</td>
<td>23</td>
</tr>
<tr>
<td>2.1.5</td>
<td>Pseudopotentials</td>
<td>25</td>
</tr>
<tr>
<td>2.1.6</td>
<td>Band structure and electronic density of states</td>
<td>26</td>
</tr>
<tr>
<td>2.1.7</td>
<td>Spontaneous polarization</td>
<td>30</td>
</tr>
<tr>
<td>2.1.8</td>
<td>Magnetism</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Learning Methods</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>Data mining and ML algorithms</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Correlation Matrix</td>
<td>39</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Principal Component Analysis (PCA)</td>
<td>40</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Median Analysis</td>
<td>42</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Traditional ML algorithms</td>
<td>44</td>
</tr>
<tr>
<td>2.3.4.1</td>
<td>Regression Algorithms</td>
<td>45</td>
</tr>
<tr>
<td>2.3.4.2</td>
<td>Classification algorithms</td>
<td>47</td>
</tr>
<tr>
<td>2.3.5</td>
<td>NN algorithms</td>
<td>48</td>
</tr>
<tr>
<td>2.3.6</td>
<td>Learning Curves</td>
<td>52</td>
</tr>
</tbody>
</table>
CONTENTS

2.4 ML algorithms in the materials domain ........................................... 54

3 Human Learning .............................................................................. 57

3.1 Molecular- and polymer-based ferroelectric crystals .......................... 58

3.1.1 Ferroelectricity in PVDF-TFP ...................................................... 58

3.1.1.1 Introduction ........................................................................ 59

3.1.1.2 Structural model construction .............................................. 61

3.1.1.3 Computational techniques .................................................. 64

3.1.1.4 Discussion ........................................................................ 65

3.1.1.5 Summary .......................................................................... 72

3.1.2 Ferroelectricity in DIPA-X ........................................................... 73

3.1.2.1 Introduction ........................................................................ 74

3.1.2.2 Crystal symmetry and structural models ............................... 76

3.1.2.3 Computational details ........................................................ 80

3.1.2.4 Discussion ........................................................................ 84

3.1.2.5 Influence of unit dipole canting on polarization .................... 91

3.1.2.6 Summary .......................................................................... 93

3.2 Derivatives of BFO ....................................................................... 95

3.2.1 La, Sr-doped BFO .................................................................. 97

3.2.1.1 Introduction ........................................................................ 97

3.2.1.2 Computational techniques .................................................. 99

3.2.1.3 Discussion ........................................................................ 101

3.2.1.4 Summary .......................................................................... 109

3.2.2 Bi$_2$FeMnO$_6$: a double perovskite ......................................... 111

3.2.2.1 Introduction ........................................................................ 112

3.2.2.2 Computational techniques .................................................. 113

3.2.2.3 Discussion ........................................................................ 115

3.2.2.4 Summary .......................................................................... 118

3.3 Chapter Summary .......................................................................... 120

4 Traditional ML-algorithms based learning ...................................... 122
CONTENTS

4.1 Crystallization propensity of APIs ........................................... 123
   4.1.1 Introduction ......................................................... 123
   4.1.2 Methods ........................................................... 128
      4.1.2.1 Data acquisition and curation .......................... 128
      4.1.2.2 Algorithms, descriptors and model building ....... 129
   4.1.3 Results and Discussion .......................................... 131
      4.1.3.1 API property trend analysis .............................. 131
      4.1.3.2 API only models ............................................ 133
      4.1.3.3 API + solvent models .................................... 140
   4.1.4 Conclusions ....................................................... 146
4.2 Magnetism in Actinide Systems ............................................. 147
   4.2.1 Introduction ......................................................... 147
   4.2.2 Methods ........................................................... 152
      4.2.2.1 Datasets ...................................................... 152
      4.2.2.2 Descriptors ................................................ 158
      4.2.2.3 Data Analytics .............................................. 159
      4.2.2.4 Algorithms ................................................ 162
      4.2.2.5 ML model development and validation ............... 162
   4.2.3 Results and Discussion .......................................... 164
   4.2.4 Conclusions ....................................................... 177
4.3 Organic Ferroelectrics ................................................... 178
   4.3.1 Introduction ......................................................... 178
   4.3.2 Methods ........................................................... 183
      4.3.2.1 Data Compilation and Curation .......................... 183
      4.3.2.2 Descriptors and ML models ............................... 183
      4.3.2.3 Data Analysis .............................................. 187
   4.3.3 Results and Discussion .......................................... 190
      4.3.3.1 Shortlisting potential organic ferroelectrics ....... 192
   4.3.4 Summary ............................................................ 198
4.4 Chapter Summary .......................................................... 198
<table>
<thead>
<tr>
<th>5 Lessons learned and concluding remarks</th>
<th>202</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Appendix</td>
<td>205</td>
</tr>
<tr>
<td>6.1 Specifications of ML Models</td>
<td>205</td>
</tr>
<tr>
<td>6.1.1 Hyperparameters used for ML models constructed in Chapter 4.1</td>
<td>205</td>
</tr>
<tr>
<td>6.1.2 Hyperparameters used for ML models constructed in Chapter 4.2</td>
<td>206</td>
</tr>
<tr>
<td>6.1.3 Structures of molecular compounds</td>
<td>206</td>
</tr>
<tr>
<td>References</td>
<td>211</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The growing availability of databases collating theoretical and experimental materials data across disciplines has opened new avenues for employing data-driven approaches utilizing *machine learning (ML)* algorithms to establish yet-unknown structure-property-performance relationships that may exist within a given family or class of materials [1]. Materials science and related fields are experiencing a paradigm shift towards using either already existing or newly generated *aggregated* knowledge to accelerate the discoveries of novel materials and advanced material functionalities. Multiple successful examples exist for technological applications in the fields of energy [2–11], catalysis [12] and photovoltaics [13], for pharmaceutical applications in drug design and reaction mechanisms mapping [14–18], as well as in advancing fundamental knowledge of materials behavior, including magnetism [19], ferroelectricity [20] and superconductivity [21]. All such applications utilize various ML algorithms in combination with data mining tech-
1. Introduction

Figure 1.1: A representative diagram showing fundamental steps to construct a ML framework. The primary components of such study includes data mining: compilation of simulations- or experiments-based relevant datasets, selection of list of descriptors followed by choice of ML algorithms to develop predictive models. Generation of models also include selection of most relevant descriptors and algorithms, optimization of model performance to predict endpoints with reasonable accuracy.

While there are several types of ML algorithms in use, a few of most popular ones are briefly discussed below. Traditional supervised learning algorithms include major classification and regression [22] that are mainly applied to small datasets, containing
hundreds of data points. On the other hand, deep neural networks (NN) are usually utilized when large datasets, containing thousands to millions of data points, are available. Variations of NNs incorporate, but are not limited to architectures built using single and multi-layers perceptrons, feed forward, radial basis, convolutional, general adversarial networks. These algorithms are also discussed in more details in Chapter 2.

In supervised machine learning, the response or output variable is represented by a function of the given input variables. The mapping between input and output data is well-defined, which allows for predictions to be made utilizing new input data. ML models employing this approach are optimized based on the bias-variance trade-off and model complexity. Bias is the difference between the average prediction of ML model and the true value which the model is trying to predict. Variance is defined as the variability of model prediction for a given data point or a value which tells us spread of our data. The error due to bias is taken as the difference between the expected (or average) prediction of model and the true or correct value that the model is trying to predict. The variability of a model prediction for a given data point is referred to as the error due to variance. If the model is too simple and has very few hyperparameters, then it may have high bias and low variance. If the model has large number of parameters then it will have high variance and low bias. Hence, a good supervised model tries to achieve a balance between bias and variance that eventually minimizes the total error. In other words, the bias-variance tradeoff will find a reasonable balance such that the model will not overfit or underfit data.
1. Introduction

Unlike the traditional models, deep NN-based ones are built solely on the input data, which constitutes \textit{unsupervised learning}. The focus of such learning mode is to establish patterns directly in the input data distributions – commonly using clustering and density estimation methods – which usually necessitates much bigger datasets (thousands to millions of data records). Therefore, the choice of the specific learning model and the associated algorithms is usually contingent not only on the nature of the specific problem being investigated (and the target to be predicted) but also on the availability and abundance / scarcity of the relevant data.

Various examples of utilization of both traditional and NN-based ML algorithms are already available in the domain of materials science and include predictions of new materials [23–26] and evaluations of numerous material properties [27–30]. All of these studies aim to bridge the gap between experiment and theory such that the predictions obtained using ML models built on theoretical datasets are comparable to that reported via experiments. In addition, these investigations also promote a systematic data-intensive research approach to either study fundamentals of material properties or accelerate novel materials discovery with various applications. In the case of material properties prediction, typical studies include, but are not limited to, estimations of electronic band gaps and linear elastic properties (such as individual elastic moduli). The former are important for applications in light-emitting diodes (LEDs), photovoltaics, scintillators, or transistors, while the latter inform such characteristics as hardness, bonding strength. It is noteworthy that datasets usually used in such investigations [31–38] are primarily ob-
tained by computationally expensive density function theory (DFT) simulations, rather than experimental data. Naturally, the accuracy of DFT simulations and the quality of experimental measurements are dependent on a variety of different factors and, therefore, datasets assembled solely from computations or experiments cannot be used without further extensive data curation and validation. Ideally, most realistic physical models can be generated with ML-based frameworks aggregating both experimental and computational data, parts of which can be used for mutual cross-validation. Development of such frameworks is the first goal of this dissertation, with some results presented and discussed in Chapter 4.

Furthermore, for most investigations mentioned above, the structure-property relationships that these studies aimed to elucidate, have been known already or intuitively apparent [39–41]. In this dissertation, we attempt to target more complex (but still quite generic!) material properties or behavior — such as magnetism and magnetic ordering, ferroelectricity and crystallization propensity — where possible connections between accessible system features / descriptors and predicted endpoints are much more obscure. Hence, the second goal of this dissertation is to explore the utility of data-driven approaches for covering such knowledge gaps and use them to identify primary driving factors that underpin complex materials behavior for each studied case in a comprehensive manner.

Specifically, the first case, described below in Section 4.1, involves construction of a generalized framework aggregating experimental data to evaluate crystallization propen-
1. Introduction

sity of small organic molecules, as compiled by Pfizer Inc. Crystallization [42, 43] is an
teresting procedure for the purification and isolation of specialty chemicals. For new
molecular compounds, understanding their inherent tendency to crystallize is a key step
for developing efficient industrial processes for producing them in their solid form. To
date, understanding the crystallization propensity of molecular solids has been primar-
ily driven by empirical approaches. These studies show that crystallization may po-
tentially be influenced by a wide variety of processing parameters, including those of
structural, thermal, chemical and kinetic kind, but the precise nature of such connections
remains elusive. Unfortunately, detailed trial-and-error studies that evaluate crystalliza-
tion propensities for a diverse range of compounds have been lacking because they are
expensive, time-consuming and inefficient, but more importantly failed experiments are
rarely reported in the literature [44–51]. The absence of data with failed crystallization
outcomes creates an inherent bias in the dataset. As a result, a model developed based
on a dataset with mostly successful crystallization outcomes is only capable of predict-
ing a crystallization success scenario but will not perform well in the case of molecules
that are difficult to crystallize. Computational data-driven approaches utilizing aggre-
gated historical data processed by ML methods could be useful to a number of industries
that depend on successful crystallizations of new molecular entities. A number of ma-
chine learning algorithms have already been developed to predict crystallization rates
[43–49, 51]. Although, most of these models were constructed based purely on theoreti-
cal datasets, while operational conditions for crystallization experiments — such as, e.g.,
working temperature, solvents used, presence or absence of crystal seeds, sample degra-
dation, etc. — were not taken into consideration, which limited the predictive accuracy
of such models to \(\sim 30\%\) RMSE. Here, we have used a regression-based ML training ap-
proach for constructing models that predict crystallization propensity of pharmaceutical
compounds from a combination of readily available molecular descriptors and historical
processing data (obtained from Pfizer Inc. and containing records for a variety of experi-
mental factors, including information on failed crystallization experiments). We observed
that, while models built solely on descriptors of pharmaceutical molecules can predict the
endpoint with 70\% RMSE (i.e., have the same accuracy as the models developed before),
inclusion of solvent information improves this accuracy to 80\%. Beyond the insertion
of solvents, experimental factors, such as the presence of impurities and/or degradants,
were found to have the greatest influence on the model accuracy.

Due to the the materials-agnostic nature of ML algorithms, we were able to develop
similar frameworks for probing other interesting material functionalities, such as mag-
netism in actinides described in Section 4.2 and emergence of ferroelectricity in molecular
and polymeric crystals described in Section 4.3. For both of these projects, datasets based
on computational and experimental reports were compiled, curated and then used to con-
struct performance-optimized traditional ML-algorithm based models for predicting the
appropriate properties of interest, such as, e.g., magnetic moment arrangements or values
of spontaneous ferroelectric polarization. We found that supervised ML frameworks em-
ploying regression and classification algorithms can be utilized to predict magnetic mo-
ment size and ordering in uranium-based binary compounds. Descriptors representing compound structural parameters and cation f-subshell occupation numbers were identified as most important for making accurate predictions. The moment size (spin and orbit) can be predicted with \( \sim 85\% \) accuracy, whereas the magnetic ordering type can be predicted with an accuracy of 76%.

For evaluating the ferroelectric properties of organic compounds, we have collected available reports (both computational and experimental) on most of already known molecular and polymeric crystal ferroelectrics, which constitutes \( \sim 80 \) entries and is significantly less than an optimal number of data records needed to build and validate a reliable ML model. To overcome this restriction, we utilized a Gaussian-based data generation algorithm to artificially increase the number of training data points in the dataset, while testing the model only on the real compound data. We identified a combination of structural and molecular descriptors that are associated with the presence of ferroelectricity (in an organic crystalline compound) and developed a routine for predicting the associated spontaneous polarization with average prediction accuracy of \( \sim 80\% \). The same ML model was then used to evaluate the propensity for exhibiting ferroelectric properties in a large group of organic materials found in the Cambridge Crystallographic Data Center (CSD) [52] of molecular systems.

In general, multifunctional ferroelectrics are widely employed in a variety of technological applications, including acoustic, nonlinear-optic and electromechanical devices — e.g., sensors, actuators and transducers — as well as in pyroelectric arrays, nonvolatile
memories, and high-$k$ dielectric components for microelectronics. The perovskite-type $\text{ABO}_3$ metal-oxide ceramics, such as $\text{BaTiO}_3$, $\text{PbTiO}_3$, $\text{KNbO}_3$, or solid solutions, such as $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$ (PZT), usually exhibit the strongest polar properties, i.e., spontaneous polarization and piezoelectric, and pyroelectric response. However, the usefulness of oxide ceramics is limited by their substantial weight, brittleness, toxicity (e.g., in lead-containing compounds) and substantial costs of device manufacturing [53, 54]. The advantage of molecular or polymer based ferroelectrics is that, while displaying more modest polar properties, they are lightweight, flexible and environmentally friendly [55–60]. Furthermore, a range of practical and inexpensive prescriptions is available for their synthesis and processing, as either bulk materials, or nanostructures. Nonetheless, in order to fully realize the potential of polar polymer and molecular based materials for modern technological applications, it is paramount to acquire detailed molecular-level understanding of mechanisms governing the emergence of ferroelectricity in them, which is the main goal of the project described in Section 4.3.

In the course of carrying out the described projects, we have also learned that expensive computations (such as those requiring the usage of DFT) needed for the determination of materials properties related to magnetism and ferroelectricity can be avoided, to a certain extent, by employing data-driven ML approaches instead. Quantification of the relative importance of features controlling such complex materials functionalities also helps facilitate renewed interest in accelerating the design of novel compounds for a wide variety of technological applications.
Finally, in order to develop our general understanding of functional (ferroelectric and magnetic) materials at the electronic level, we have conducted a number of ‘conventional’ studies of their properties utilizing DFT. These projects, called ‘Human Learning’ in what follows, nonetheless provided us with useful initial insights into the behavior of such materials at the atomistic scale that were later utilized for the development of ML and data-driven workflows. In particular, we have investigated the electronic and polar properties of two organic-ferroelectric families of compounds, namely, novel polyvinylidene difluoride (PVDF) based mixed polymer crystals incorporating 2,3,3,3-tetrafluoropropene (TFP, -CH$_2$-CF(CF$_3$)-) and organic crystals formed by combining diisopropylammonium (DIPA) molecules and different halide counter ions (X), as discussed in Section 3.1.1. These studies led to detailed molecular-level understanding of mechanisms governing the emergence of spontaneous polarization in such functional materials.

We have also explored the effects of dopants and vacancies on the electronic, magnetic and optical properties of two Bismuth Ferrite (BFO)-derived systems: (La, Sr) doped BFO and double perovskite Bi$_2$FeMnO$_6$ (BFMO). These studies revealed an enhancement of polarization in BiLaFe$_2$O$_6$, as compared to pristine BFO. The local magnetic moment of Fe atoms decreases upon dopant substitution and its magnitude is dependent on the distance between the Fe and dopant atoms. The investigation of oxygen vacancy effects of BFMO demonstrated that a strong on-site Hubbard interaction is critical for the gap opening in pristine BFMO. The average magnetization decreases with the increase of oxygen vacancy concentration. From the calculated band structure and optical conductivity,
an insulator-metal transition or crossover was identified in BFMO. To sum up the main findings of the studies, we provide summaries of both Human Learning and Machine Learning investigations below.

**Human Learning Summary:** PVDF-based mixed polymer crystals incorporating TFP and organic crystals formed by combining DIPA molecules and different halide counter ions (X) were investigated utilizing first principles theory in conjunction with Wannier function charge center analysis. These studies led to detailed molecular-level understanding of mechanisms governing the emergence of spontaneous polarization in such functional materials. While in the PVDF-TFP material templates, the electronegativity difference between H and F atoms play a key role for generating polar systems, such properties exhibited by DIPA-X crystals are driven by structural arrangements of halide and ammonium ions. For the two BFO-derived systems, (La, Sr) doped BFO and double perovskite BFMO, the effects of dopants and vacancy were explored to quantify the effects of such defects on the electronic, magnetic and optical properties. Our study revealed an enhancement in polarization in BiLaFe$_2$O$_6$ as compared to pristine BFO. The local magnetic moment of Fe atoms decreases upon dopant substitution and its magnitude is dependent on the distance between the Fe and dopant atoms. The investigation of oxygen vacancy effects of BFMO demonstrated that a strong on-site Hubbard interaction is critical for the gap opening in a pristine BFMO. The average magnetization decreases with the increase of oxygen vacancy concentration. From the calculated band structure and optical conductivity, an insulator-metal transition or crossover was identified in BFMO.
Machine Learning Summary: Employing the insights obtained from the Human Learning studies, we have investigated three different cases that involve predicting advanced functionalities of materials systems — molecular crystallization propensity, magnetic moment strength and arrangements, and predisposition for ferroelectricity. The algorithms used in the projects as discussed in Chap. 4 are various regression-based algorithms [22] such as Linear (LR), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RFR), Kernel Ridge (KRR), Support Vector Machine (SVMR) regressions and Random Forest classification (RFC). Each target functionality is connected to a variety of applications, including those in pharmaceutical sciences or modern electronic devices. We demonstrate that appropriate data-driven techniques can facilitate our understanding of key aspects of such complex processes: e.g., in the case of crystallization of pharmaceutical molecules, providing valuable insights for developing better medicaments with extended shelf life. Different molecular descriptor combinations, the sizes of training sets used, as well as a variety of experimental factors were taken into consideration in that study and their influence on the predictive accuracy of the developed models was evaluated. We found that, while the ML models built solely on active pharmaceutical ingredients (API) can predict the endpoint with 30% RMSE, inclusion of solvents into such models has improved this accuracy to 20%. Beyond the insertion of solvents, the presence of impurities and/or degradants has the greatest influence on model accuracy.

As the next case, the origins of magnetic behavior of actinide systems were studied on the electronic level. We established structure-property links for these systems by assem-
bling and mining two datasets that aggregate, respectively, the results of DFT simulations and experimental measurements for the families of uranium- and neptunium-based binary compounds. Various regression algorithms were used to identify correlations among accessible attributes (features or descriptors) of the material systems and predict their cation magnetic moments and general forms of magnetic ordering. The moment size (spin and orbit) was predicted with ∼85% accuracy, whereas the models to predict ordering had accuracy of 76%.

In the last investigation, we developed a ML-based framework to predict the magnitude of spontaneous polarization in organic compounds by collecting and analyzing reports (both computational and experimental) on existing organic ferroelectric materials. During the ML model construction, a specific combination of structural and molecular descriptors was identified as the most important for predicting ferroelectric polarization, with an average endpoint prediction accuracy of ∼80%. The developed model was then utilized to conduct a (computationally inexpensive) screening of a large variety of molecular crystals to shortlist materials that may potentially exhibit ferroelectricity.

In conclusion, we demonstrate the remarkable flexibility and ‘material agnostic character’ of data-driven and ML techniques in their utilization for developing better insights into complex materials behavior and specifically elucidating the mechanisms governing sophisticated functionalities and accelerating rational search for and design of novel materials for advanced applications.
The rest of the dissertation is organized as follows:

Chapter 2 presents the details of the utilized generic computational methodologies, including both DFT and ML-based approaches. The discussion of DFT-based techniques, provided in Section 2.1, includes band structure and electronic density of states (EDOS) computations, as well as methods for evaluating the spontaneous polarization and magnetization. Various forms of learning methods, several data mining and ML algorithms, along with their relevance for applications in materials science are discussed in Section 2.2. We note that even though many of these techniques are common for the discussed projects, additional details relevant to each specific study are listed in respective chapters.

Chapter 3 reports the ‘Human Learning’ investigations of electronic and polar properties of novel PVDF-TFP and DIPA-X ferroelectric crystal families (Section 3.1), as well as electronic, polar, magnetic and optical properties of the BFO-derived perovskite structures (Section 3.2).

In Chapter 4, the detailed results of ML-based studies of crystallization propensity of small organic molecules, magnetism occurring in uranium-based compounds and ferroelectricity observed in organic systems are presented in Sections 4.1, 4.2 and 4.3, respectively.

General conclusions of the dissertation are given in Chapter 5.
Chapter 2

Computational Methods

This chapter aggregates the descriptions of generic computational techniques used to obtain results discussed in the following chapters. A brief review of the DFT methodology, including Kohn-Sham equations, approximations used to represent the electron charge density and pseudopotentials, is presented in Section 2.1 along with examples of typical calculations of electronic band structure, EDOS, spontaneous polarization and magnetic properties. In the same venue, Section 2.2 briefly discusses various types of learning methods, data mining techniques and ML algorithms, i.e., the computational and -statistical tools that are used to build ML-based frameworks. Any specific details of calculations utilized in individual projects that are described later are provided in the appropriate sections of Chapters 3 and 4.
2.1 Density Functional Theory

2.1.1 General Considerations

Electronic properties of a crystal or molecule can be evaluated from ‘first principles’ by solving a *Schrödinger equation*. A time-independent, nonrelativistic version of this equation has the following form:

\[ \mathcal{H}\Psi = E\Psi, \]  

(2.1)

where \( \Psi(\mathbf{R}_I, \mathbf{r}_i) \) is a many-electron wave function, with \( \mathbf{R}_I \) and \( \mathbf{r}_i \) representing positions of all nuclei \( I \) and all electrons \( i \), respectively. By operating on the \( \Psi(\mathbf{R}_I, \mathbf{r}_i) \) with Hamiltonian operator \( \mathcal{H} \), total energy of the system \( E \) can be obtained.

The Hamiltonian operator for a system consisting of interacting electrons and nuclei can be expressed as sum of the following operators, each representing different interactions within the system:

\[ \mathcal{H}(\mathbf{R}_I, \mathbf{r}_i) = T_e(\mathbf{r}_i) + T_N(\mathbf{R}_I) + V_{ee}(\mathbf{r}_i, \mathbf{r}_j) + V_{eN}(\mathbf{r}_i, \mathbf{R}_I) + V_{NN}(\mathbf{R}_I, \mathbf{R}_J). \]  

(2.2)

Here \( T_e \) and \( T_N \) are kinetic energy operators for the electrons and nuclei, respectively, and \( V_{ee}, V_{eN} \) and \( V_{NN} \) are electrostatic energy operators for the electron-electron, electron-
nuclear and nuclear-nuclear interactions. Each operator is described in detail below:

\[ T_e = -\sum_i \frac{\hbar^2}{2m_e} \nabla_i^2, \quad (2.3) \]

\[ T_N = -\sum_i \frac{\hbar^2}{2M_I} \nabla_i^2, \quad (2.4) \]

\[ V_{ee} = \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|r_i - r_j|}, \quad (2.5) \]

\[ V_{eN} = -\sum_{i,I} \frac{Z_I e^2}{|r_i - R_I|}, \quad (2.6) \]

\[ V_{NN} = \frac{1}{2} \sum_{i \neq j} \frac{Z_I Z_J e^2}{|R_I - R_J|}. \quad (2.7) \]

Here, \( Z_I \) and \( M_I \) are the charge and mass of the nucleus \( I \), and \( m_e \) and \( e \) are the mass and charge of an electron.

Due to the much larger mass of the nuclei (3 to 5 orders of magnitude larger than \( m_e \)), it is reasonable to assume that electrons can adjust almost instantaneously to the change in positions of the nuclei. As a result, the kinetic energy term \( T_N \) in Eq. (2.2) can be decoupled from the rest of the energy terms in the same equation. The process of decoupling of the nuclei and electron degrees of freedom into separate equations is called the Born-Oppenheimer or adiabatic approximation. The lowest energy configuration or ground state of the electron system (with static nuclei) is then obtained by solving an equation involving an electron Hamiltonian \( \mathcal{H}_e \):

\[ \mathcal{H}_e(R_I, r_i) = T_e(r_i) + V_{ee}(r_i, r_j) + V_{eN}(r_i, R_I) + V_{NN}(R_I, R_J). \quad (2.8) \]

Unfortunately, even in this approximation, as described by Eq. (2.8), the problem of eval-
2. Computational Methods

Evaluating the electronic properties of the system is still too difficult to be solved efficiently for most practical materials [61].

2.1.2 Kohn-Sham Equations

DFT was developed to avoid using complicated formulations that involve the many-particle wave function $\Psi(R_i, r_i)$, instead expressing the necessary system observables as functions of its electron charge density only. This approach, representing a reasonable trade-off between computational effort and achievable accuracy, has proven to be very popular as an efficient tool for predicting materials properties [62–64]. A non-exhaustive list of such properties, that could be evaluated for periodic crystals or aperiodic structures like molecules and clusters, includes equilibrium atomic positions, distribution of electronic energies, vibration frequencies, ionization potentials, electron affinities, binding and cohesive energies and magnetic properties.

As follows from two theorems listed below, DFT allows one to determine the electronic ground-state of a system by solving Kohn-Sham equations:

1. The ground-state solution of the Schrödinger equation is a unique function of the electron density. In other words, there exists a one-to-one mapping between the system ground-state wave function and electron charge density $\rho(r)$. Furthermore, the ground-state energy $E$ can be expressed as $E[\rho(r)]$.

   An important consequence of this theorem is that the knowledge of $\rho(r)$ uniquely
determines all of the system ground state properties, including the energy and wave function. This result is extremely important because the equations describing the system can now be solved by finding a function of three spatial variables, i.e., the electron charge density $\rho(\mathbf{r})$, instead of the many electron wave function, which is a function of many variables. For example, the problem of finding the ground state of a nanoparticle consisting of 100 Pd atoms is reduced from more than 23,000 dimensions to a problem with only 3 dimensions [61]. Nevertheless, the functional form of the $E[\rho(\mathbf{r})]$ dependence is still unknown, but some of its properties are described by the second theorem.

2. The electron charge density that minimizes the energy of the overall functional is the true electron charge density corresponding to the full solution of the Schrödinger equation. Therefore, if the functional form of $E[\rho(\mathbf{r})]$ is known, the ground state of the system can be obtained by minimizing this functional.

Thus, the DFT approach replaces the more elaborate picture of a many-electron interacting system by a picture where a ‘mean field’ system with independent electrons is considered that nonetheless faithfully reproduces the effects of many-electron interactions.
Figure 2.1: Flowchart showing individual steps in the self-consistent solution of the Kohn-Sham equations in order to obtain ground state energy and optimized crystal structure of a system.
For many-electron systems:

\[ E[\{\psi_i\}] = E_{\text{known}}[\{\psi_i\}] + E_{\text{XC}}[\{\psi_i\}], \quad (2.9) \]

\[
E_{\text{known}}[\{\psi_i\}] = -\frac{\hbar^2}{m_e} \sum_i \int \psi_i^* \nabla^2 \psi_i d^3r + \int V(r) \rho(r) d^3r \\
+ \frac{e^2}{2} \int \int \frac{\rho(r) \rho(r')}{r - r'} d^3r d^3r' + E_{\text{XC}}[\{\psi_i\}].
\]

There exists a unique one-to-one mapping is established between a system containing N interacting electrons with charge density \( \rho(r) \) moving in an external potential \( V(r) \) and a fictitious system of N non-interacting electrons also with the same change density \( \rho(r) \). The Schrödinger equation is effectively replaced by the set of Kohn-Sham equations formulated for single-electron wave functions \( \psi_i(r) \). The ‘known’ energy terms include electron kinetic energies, electron-nuclei Coulomb interactions, electron-electron Coulomb interactions and nucleus-nucleus Coulomb interactions. The remaining term in the total energy description, \( E_{\text{XC}}[\{\psi_i\}] \), or the exchange–correlation energy term, aggregates all the other quantum mechanical effects in the system that are not precisely known. This last term has to be approximated in some way to allow the Kohn-Sham equations

\[
\left[ -\frac{\hbar^2}{2m_e} \nabla^2 + V_{eN}(r) + V_{ee}(r) + V_{\text{XC}}(r) \right] \psi_i(r) = \epsilon_i \psi_i(r)
\]

(2.11)

to be solved, e.g., iteratively, as shown in Fig. 2.1.
2.1.3 Further Approximations

With the exception of the Born-Oppenheimer or adiabatic approximation, the derivations presented in the previous section are exact, meaning that no other specific approximations have been introduced yet. However, in order to solve the Kohn-Sham equations for practical systems, further simplifications, including those specifying the form of the exchange-correlation potential, are necessary to make numerical calculations tractable.

**Exchange-correlation energy:** DFT methodology allows us to obtain information about the ground-state of the system by minimizing the Kohn-Sham energy functional through variation of the electron charge density function. This can be achieved by iteratively solving a set of single-electron wave function equations (2.11). In order for the DFT calculations to be useful for practical purposes, some approximation should be adopted for the form of the exchange-correlation potential $E_{XC}[\{\psi_i\}]$.

For example, one of the earliest and simplest approximations is the so-called *Local Density Approximation* (LDA) [65]. In fact, LDA is the exact exchange-correlation potential for the homogeneous electron gas — i.e., a system whose electron charge density is constant throughout the space, $\rho(\mathbf{r}) = \text{constant}$ [61]. In general, the LDA functional is one of the most popular and widely used ones [66] for the DFT calculations. In this dissertation, most calculations (except those involving BiFeO$_3$ systems that require more elaborate treatment) are done using LDA. Within LDA, the exchange-correlation energy of the system is assumed to correspond locally to that of the uniform electron gas with
same electron density, i.e.,

\[ V_{XC}^{\text{LDA}} = V_{XC}^{\text{LDA}}[\rho(r)] \tag{2.12} \]

In real systems, the electron charge density is not homogeneous, and therefore the LDA exchange-correlation potential is considered to be a rather simple approximation (but a very useful one nonetheless). Besides LDA, there are several other standard and well known exchange-correlation potentials. For example, generalized gradient approximation (GGA) is expressed as a function of both electron density and its gradient [67–71].

\[ V_{XC}^{\text{GGA}} = V_{XC}^{\text{GGA}}[\rho(r), |\nabla\rho(r)|] \tag{2.13} \]

As of now, there are over 200 exchange-correlation functionals and a curious reader can consult articles such as Ref. [72] for a comprehensive review.

2.1.4 System periodicity

Applying DFT formalism to bulk systems involves yet another complication due to the explosion of the number of nuclei and electrons in the system. Crystallographic periodicity is exploited for the bulk materials, as they are typically modeled by atoms occupying equilibrium positions in a unit cell representative of the crystal symmetry that is repeated throughout space by periodic translations. Therefore, all the potential energy terms present in the Kohn-Sham equations must have the same periodicity as that of the
crystal in question, i.e.,

\[ V(r + \mathbf{R}) = V(r). \]  

(2.14)

Here \( \mathbf{R} \) is the real space lattice translation vector that in 3D can be represented as a linear combination of three lattice vectors, \( \mathbf{a}, \mathbf{b} \) and \( \mathbf{c} \):

\[ \mathbf{R} = n_1 \mathbf{a} + n_2 \mathbf{b} + n_3 \mathbf{c}. \]  

(2.15)

where \( n_j \) is an integer.

The solutions of the Schrödinger equation for such periodic systems should also satisfy the periodic boundary conditions and, according to Bloch’s theorem [73], the appropriate electronic wave functions can be expressed as

\[ \psi_k(\mathbf{r}) = e^{i\mathbf{k} \cdot \mathbf{r}} u_k(\mathbf{r}), \]  

(2.16)

where \( u_k(\mathbf{r}) \) is lattice periodic. The Schrödinger equation can then be solved independently at each value of \( \mathbf{k} \), which is a vector of the reciprocal space of the system represented by its Brillouin Zone (BZ) [73].

**k-point sampling:** In practical DFT calculations, observable system properties are usually averaged over the BZ by evaluating the following integral:

\[ \langle g \rangle = \frac{1}{V_{\text{BZ}}} \int_{\text{BZ}} d\mathbf{k} g(\mathbf{k}) \]  

(2.17)

Numerically, the integration over the reciprocal space has to be discretized as a summation over the grid of \( k \)-points. Several standard approaches to sample \( k \)-points in the reciprocal space are available. For example, Monkhorst-Pack (MP) sampling [74] is a careful
selection of $k$-points on a 3D grid to ensure the accuracy and efficiency of the calculations. Usually, the $k$-point grid density has to be carefully optimized for specific systems to provide the most accurate results in reasonable time, as tested by the convergence of total energy with respect to the grid density.

### 2.1.5 Pseudopotentials

Until now, in order to solve Kohn-Sham equations, as shown in Eq. (2.11), the electronic wave functions were assumed to be expressed in terms of, e.g., plane wave basis functions, and the solutions were obtained in reciprocal space. Since valence electrons dominate the chemical bonding and interatomic interactions in materials, it is possible to avoid treating all the core electrons explicitly by using pseudopotentials [61]. Typically, pseudopotentials have to be developed separately for each atomic type. They also have to be thoroughly tested for transferability, i.e., their accuracy in different chemical environments, e.g., such as a drastic change of coordination.

In general, DFT within various approximations (LDA, GGA, meta-GGA, hybrid GGA, etc.) is still a powerful tool, widely used in materials modeling due to its transferability and simplicity, and often provides a starting point for more accurate investigations of materials behavior. We have used projector augmented wave (PAW) method in our projects for generalizing pseudopotentials (PBE [71] and ultrasoft [75]) to conduct DFT calculations with reasonable efficiency. The rest of this section focuses on the details of the phys-
2. Computational Methods

(a) Structural model of a rhombohedral perovskite crystal system with 10-atoms unit cell. (b) A representative band structure corresponding to this system in [-4,4] eV energy range as computed along high-symmetry BZ directions for a rhombohedral crystal system (F-Γ-Z-L-F). Energy is measured with respect to the system Fermi energy level, which is taken as zero.

Figure 2.2: (a) Structural model of a rhombohedral perovskite crystal system with 10-atoms unit cell. (b) A representative band structure corresponding to this system in [-4,4] eV energy range as computed along high-symmetry BZ directions for a rhombohedral crystal system (F-Γ-Z-L-F). Energy is measured with respect to the system Fermi energy level, which is taken as zero.

2.1.6 Band structure and electronic density of states

Electronic band structure of a periodic system provides the allowed electronic energy levels throughout the BZ. It is usually represented by 2D plots of the dispersion of the energies of the single-electron Kohn-Sham wavefunctions, each characterized by its band index $n$, along a certain direction of the Bloch vector $\mathbf{k}$. A sample band structure of a
rhombohedral (perovskite) crystal system is shown in Fig. 2.2. According to the Bloch’s theorem, the energy eigenstates for an electron in a crystal can be written as Bloch waves where the wavefunctions are of the following form:

$$\psi_k(r) = e^{i\mathbf{k} \cdot \mathbf{r}} u_k(r),$$  \hspace{1cm} (2.18)

where \( r \) is position, \( \psi_k(r) \) is the Bloch wave, \( u_k(r) \) is a periodic function with same periodicity as the crystal, \( \mathbf{k} \) is the wave vector [same as Eq. (2.16)]. The Bloch vector is an element of the reciprocal space and these vectors are restricted to the first BZ. A corresponding energy band \( \epsilon_n(k) \) may be perceived as continuous representation for a collection of eigenvalues of an equation, such as Eq. (2.11), that is being solved for the single-electron wavefunction \( \psi_{n,k}(r) \) as the wave vector \( \mathbf{k} \) is changing along some trajectory (usually a high-symmetry line or other element) within the BZ.

By Pauli’s principle, each state with a given \( n \) and \( k \) can accommodate no more than two electrons (of opposite spin) and consequently each band can have two electrons per primitive unit cell. The ground state of the system is then obtained by filling up the bands with electrons moving from lower to higher energies until all the available electrons are placed. The energy value that separates the filled bands from empty ones is called the Fermi energy \( (E_f) \) and is often used as a reference point to separate the conduction and valence bands. The band gap — in an insulating system — can then be defined as the energy difference between the lowest point of the conduction band (conduction band minimum or CBM) and the highest point of the valence band (valence band maximum or...
VBM). This gap can be direct or indirect in nature. A direct band gap is characterized by having the band edges aligned in \( k \), such that an electron can jump from the valence band to the conduction band, with the emission of a photon, without changing considerably the momentum. For the indirect band gap the band edges are not aligned and therefore both a photon and phonon are responsible for the transition of electron to conduction band.

Band structure plots are very useful for visualizing the \( k \)-dependence of the energy states and the location of the band-gap, as well as for providing insight on the possible nature of electronic transitions. However, they are usually computed along high symmetry BZ directions only and even in such cases (of high symmetry) these plots can become quite complicated for many compounds. On the other hand, density of electronic states can be sampled across the entire BZ, providing a visually simple and intuitive aggregated snapshot of the same information, i.e., how sparsely or densely the energy levels are distributed along the energy axis. EDOS is usually calculated by removing the dependence of the energy bands on the location within the BZ through a summation over all the \( k \)-points:

\[
\text{EDOS}(\epsilon) = \sum_k \omega_k \sum_i \beta(\epsilon - \epsilon_{ik}).
\]

Here \( \omega_k \) is a weight associated with the \( k \)-point and \( \beta \) is a broadening function (typically a Gaussian) used to account for the finite number of \( k \)-points used in the calculations. The amount of broadening is arbitrary and is usually chosen to provide smooth looking plots that still reproduce the important features of the EDOS(\( \epsilon \)) distribution. The location
Figure 2.3: (a) Structural model of a 5-atom unit cell of the cubic perovskite PbTiO₃ (space group Pm̅3m). (b) The associated BZ with outlined high-symmetry directions [obtained from the Bilbao Crystallographic Server [76]]. (c) Total EDOS of the system and PDOS cures for all five of the Ti d-states.

of the Fermi level on the EDOS plot is determined in the same fashion as in the case of the band structure, but could be somewhat less precise due to the broadening involved.

The projected/partial density of states (PDOS) can also be constructed to represent rel-
ative contributions of specific atoms or atomic orbitals to the total DOS. A sample EDOS plot along with a sketch of the associated BZ high-symmetry directions [76] for a 5-atom unit cell of cubic ($Pm\bar{3}m$) PbTiO$_3$ is shown Fig. 2.3 and includes both the total EDOS of the system and PDOS curves for all five of the Ti $d$-states. The latter plots show that in the case of cubic symmetry, PDOS for the Ti $d$-states is degenerate: triply for $d_{xy}$, $d_{xz}$ and $d_{yz}$ orbitals, and doubly for $d_{z^2}$ and $d_{x^2-y^2}$ orbitals.

2.1.7 Spontaneous polarization

In quantum-mechanical computations involving periodic polar systems macroscopic polarization is usually expressed as a sum of ionic ($P_{\text{ion}}$) and electronic ($P_{\text{el}}$) terms:

$$P = P_{\text{ion}} + P_{\text{el}} = \frac{1}{V} \left( e \sum_{\tau} Z_{\tau} b_{\tau} - \int_V \rho(r) \, dr \right),$$

(2.20)

where $V$ is the unit-cell volume, $Z_{\tau}$ and $b_{\tau}$ are the ionic charge and position of the $\tau$-th atom in the unit cell, and $\rho(r)$ is the cell-periodic electron-charge density. While the ionic contribution is a simple sum over point charges, the electronic one is not well defined due to the delocalized nature of $\rho(r)$, which makes the resulting $P_{\text{el}}$ dependent on the shape of the chosen unit cell. The so-called Modern Theory of Polarization, developed in the early nineties [77, 78], provides two complementary approaches for regularizing the $P_{\text{el}}$ term by representing $\rho(r)$ as a sum over occupied Bloch states $\psi_{ik}(r)$, or Wannier functions (WFs) $W_i(r)$. I.e., in paired-electron systems,

$$\rho(r) = -\frac{2e}{(2\pi)^3} \sum_{i \, \text{occ}} \int_{\text{BZ}} |\psi_{ik}(r)|^2 \, dk = -2e \sum_{i \, \text{occ}} |W_i(r)|^2.$$  

(2.21)
Adopting the former representation leads to the well-known Berry-phase expression for $P_{el}(\lambda)$, which is an inherently multivalued quantity defined modulo $2eR/V$ — the so-called polarization indetermination quantum (PIQ) — where $R$ is a translation vector of the real-space lattice [77, 78]. Parameter $\lambda$ represents a system transformation path between state $\lambda_1$, where polarization of interest in computed, and reference state $\lambda_0$, whose polarization is already known (in many instances, but not always, naturally non-polar, e.g., centrosymmetric, structures are used as reference). Since in the Berry-phase methodology it is impossible to resolve how many multiples of $2eR/V$ may get added to the value of polarization during the system evolution from $\lambda_0$ to $\lambda_1$, such calculations are usually done at multiple points $\lambda_m$ along the transformation path $\lambda$ to ensure that

$$|\Delta P_{el}^m| \equiv |P_{el}(\lambda_{m+1}) - P_{el}(\lambda_m)| \ll \frac{2eR}{V}$$

for each $m$.

On the other hand, the WF-based approach to computing polarization, while producing the same average values as the Berry-phase method, is more advantageous for evaluating the properties of modular systems and allows one to sidestep the limitations mentioned above. Adopting the WF representation for $\rho(r)$ in Eq. (2.21) results in the following expression for $P_{el}$ (in paired-electron systems):

$$P_{el} = -\frac{2e}{V} \sum_{i \text{occ}} \langle W_i | r | W_i \rangle = -\frac{2e}{V} \sum_{i \text{occ}} \langle r_i \rangle,$$

where $\langle r_i \rangle$ are the centers of charge of (doubly occupied) WFs $W_i(r)$. The nonuniqueness of the unitary transformation connecting Bloch orbitals $\psi_{ik}(r)$ and WFs $W_i(r)$ is used to obtain functions that are ‘maximally localized’ (MLWFs) according to a specific localiza-
tion criterion, with one popular choice being the minimal sum of quadratic spreads of probability distributions $|W_i(r)|^2$ [79–81].

In modular systems — such as, e.g., polymer or molecular based crystals, where each monomer or molecule can be perceived as a somewhat separate polar unit — it is possible to partition all the ionic ($\tau$) and MLWF ($i$) charge centers into groups, each belonging to a specific unit within the simulation cell. Then a simple formula approximating the dipole moment of each unit can be written as

$$D = e \sum_{\tau \in \text{unit}} Z_{\tau} b_{\tau} - 2e \sum_{i \in \text{unit}} \langle r_i \rangle.$$  (2.23)

In some instances, like, e.g., in polymer structures, MLWF charge centers shared by neighboring units can also be identified, such as charge centers located in the middle of the -C-C- bond joining two units together along the polymer backbone — which could then be claimed by both units with a weight of 1/2.

MLWF approach is also useful for straightforward visualization of any jumps in polarization value that may occur due to an emergence of multiples of $2eR/V$ during the transformation of the system. From Eq. (2.22), it can be inferred that such jumps in the value of $P_{el}$ are equivalent to jumps in positions of MLWF charge centers $\langle r_i \rangle$ as they are being translated by the unit-cell lattice vector $R$. The same considerations apply to $P_{ion}$ and ionic positions $b_\tau$ in case any of the structural units get ‘broken up’ by periodic translations of ions in the unit cell. Therefore, when the chosen system unit cell contains no broken units and all of the MLWF charge centers are translated to their expected loca-
tions, such as the vicinity of their respective ions, or interatomic bonds, the formal value of $P_{el}$ must contain none of the additional $2eR/V$ ‘indetermination quanta.’

Although the Berry-phase approach is currently adopted as a de facto standard for polarization calculations with DFT-based codes, it does possess a few limitations. Specifically, as discussed above, it provides only average characteristics of the system, such as total value of polarization within the unit cell, or, e.g., its piezoelectric constants. That is usually sufficient for ceramic crystals — e.g., ferroelectric perovskites — however, in modular systems, such as polymer and molecular crystals, it is then impossible to obtain contributions to the system total polarization from individual structural units comprising the crystal. Furthermore, for the Berry-phase approach to work it is necessary to have a straightforward structural definition of the path $\lambda$, which may not always be clear in polymer and molecular crystals. We have used both the MLWF and Berry phase computational approaches for estimating spontaneous polarizations in organic and perovskite systems, as discussed in more detail in Chapter 3.

### 2.1.8 Magnetism

Within the DFT formalism, the electronic charge density matrix can be written as a combination of the charge and magnetization density [82, 83] as described by the follow-
\[ \rho(r) = \frac{1}{2}(n(r)I + \sigma \cdot m(r)) \quad (2.24) \]

where \( \sigma \) and \( I \) represent the Pauli and Identity matrices respectively. Hence, the potential matrix can be written as:

\[ \nu'(r) = \nu(r)I + \mu_B \sigma \cdot B(r) \quad (2.25) \]

where, \( \nu'(r) \) is the spin-density matrix and \( \mu_B \) is Bohr magneton.

For systems with collinear spin configurations, such as antiferromagnetic, ferromagnetic or ferrimagnetic states, the magnetization density can be represented by a combination of the spin-up and spin-down states.

\[ \int m(r)dr = \int (\rho^+(r) - \rho^-(r))dr \quad (2.26) \]

The resulting magnetic moment as obtained using the above equation is the spin moment. The spin-moment size is estimated as the total magnetic moment size for multiferroic oxides. However, for compounds with heavier elements as found in actinides, strong spin-orbit coupling effects are observed, which results in orbital moments providing a strong contribution to the value of the total magnetic moment. In addition, in bulk systems, the magnetization can be aligned in different crystallographic directions, which gives rise to magnetocrystalline anisotropy, with associated energy differences being important system parameters in crystals with low symmetries.

In the projects discussed in later chapters, we have used Vienna \textit{ab initio} Simulation Package (VASP) [84, 85] to evaluate the magnetic properties of material systems of inter-
Figure 2.4: Basic building blocks of ML-based studies are represented in this diagram.

In the case of BFO and its derivatives, we have obtained spin moment sizes only, while for actinides, both spin and orbital moment sizes, as well as types of magnetic ordering were explored (see Chapter 3 for more details).

2.2 Learning Methods

A typical ML-based study usually consists of the following sequence of steps: data mining / acquisition, determination of data structure, data curation, selection of the appropriate ML methods to process the data and, finally, application of these to construct and then verify a meaningful predictive model of a physical phenomenon under investigation. This general sequence of data processing steps is illustrated in Fig. 2.4.
To better understand how ML methods are applicable to the domain of materials science, it is instructive to start the discussion by outlining the benefits and shortcomings of various types of learning approaches. The same arguments are also helpful for selecting data processing schemes that are appropriate for solving any specific problems of interest. In general, for any approach, the success or failure of the model — or an ‘actor’ doing the learning — depend on a variety of important factors, such as: (a) correct identification of the model or situation component to be improved, (b) any prior knowledge that the agent may have about the problem, (c) convenient representation of collected data or model components, (d) availability of feedback to the learning process. Brief descriptions of four types of learning that are commonly used in the physical sciences domain, are provided in the following paragraphs.

**Unsupervised learning:** In unsupervised learning, the agent learns patterns in the input, even when no explicit feedback is supplied. Clustering is a common example of an unsupervised task, which detects meaningful clusters among input data. For example, a regular office goer, Mr. Neogy, traveling from one area to the other in New York City, will develop a concept or percept of “good traffic days” and “bad traffic days” based on day to day experiences, without ever consulting motor vehicles department or traffic controllers for labelled examples of each.
Supervised learning: In this mode, the agent learns from a function that maps the input-output pairs from some observed instances present in the data. In Mr. Neogy’s case, inputs are percepts about traffic conditions that he has developed over time and outputs are provided by him, where he gives directions to the cab-driver. Besides the inputs provided by Mr. Neogy, here the cab-driver can also alter the outcome based on let’s say if he passes a bus or a car or a pedestrian on the road and decides to take a different route or more time. Now, Mr. Neogy’s final outcome of reaching office on time or late is a function of states such as his own perceptions as well as the cab driver’s actions like braking, accelerating or stopping distance. The output is directly available from the agent’s percepts, the cab-driver is the environment and the final outcome can be changed if either the percept or environment changes.

Semi-supervised learning: In this mode, we are given a few labelled instances and a large set of unlabelled ones. Let’s assume, we are given a task of creating a model to predict what type of coffee a person drinks on a regular basis. We can gather some data (labelled examples) by interviewing people and / or by visiting multiple coffee shops, which would be identical to supervised learning. However, in reality, some of the people interviewed may not be truthful in their responses. Furthermore, the collected data may be inaccurate for other reasons, e.g., people not knowing specific coffee types and naming different coffee brands instead. Therefore, there is not only random noise in the data, but there are systematic inaccuracies present that can only be identified by utilizing unsuper-
vised learning techniques. In other words, noise and lack of labelled instances create a continuum between supervised and unsupervised learning modes, which constitutes the domain of applicability of semi-supervised approaches.

**Reinforcement learning:** This is another hybrid mode in which the agent learns from a series of past events or reinforcements (success or failure). Here, the model gets either rewards or penalties for the actions it performs, such as searches or trials, with a goal to maximize the total reward. For example, if Mr. Neogy reaches his office on time, that gives him an indication that he did something right along the way. It then falls on the agent to decide which of the actions prior to reinforcement had the most pronounced impact on the outcome.

### 2.3 Data mining and ML algorithms

In order to construct insightful ML models, it is most important to learn as much as possible about the available data. Therefore, prior to applying any ML algorithms to a dataset, it is good practice to evaluate the data layout or structure utilizing standard data analytics techniques. These steps help to identify types of datasets, data objects, associated attributes, correlations and trends that may be present in the data, as well as eliminate any redundancies. Common datasets can be of several types, such as: records (relational, data matrix, document or transaction data), graphs and networks (social or in-
formation networks, molecular structures), ordered (videos, temporal, spatial, sequence), spatial, images and multimedia (spatial, image, video, audio). Regardless of its type, each dataset is comprised of data objects representing a particular entity. For example, a generic university database contains data objects that represent information records for students, professors, staff, courses being taught, etc. Data objects are best described by their attributes (features or descriptors). Depending on the dataset type, these attributes can be of nominal categories (i.e., simply naming things), binary, ordinal (grades), numeric, interval- and ratio-scaled. In many instances, such attributes also fall into groups that have distinct attributes, like being discrete (e.g., binary) or continuous (e.g., numeric) in nature. Once key components of the dataset of interest are established, multiple data analytics or summarization techniques [86] can be used to evaluate the properties of the data and highlight entries that can be treated as outliers or noise. These techniques are briefly discussed in the following sections and utilized in projects discussed in Chapter 4.

2.3.1 Correlation Matrix

Correlation matrix and coefficient are used to identify any correlations that may be present among variables in a dataset. A widely used correlation method called Pearson’s correlation [87] can help find linear correlations between two data attributes using the
following formula: For a given paired data: \([ (x_1, y_1), \ldots, (x_n, y_n) ]\) with \(n\) pairs,

\[
 r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{2.27}
\]

where \(n\) is sample size, \(x_i, y_i\) are individual data points indexed with \(i\) and \(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i\) are the sample means. This coefficient varies between -1 and 1, with both of the extremum values representing the highest correlations. While +1 corresponds to the highest positive linear correlation between two specific variables, -1 represents the highest anti-correlation. In other words, if the coefficient is 1, for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. On the other hand, if the coefficient is -1, for every positive increase in one variable, there is a negative increase or decrease of a fixed proportion in the other. The coefficient value of zero represents no linear correlation between two variables.

While this correlation analysis is extremely helpful in identifying linear relationships between variables present in a dataset, it does not provide any specific information on how these variables may be related. Another drawback of this technique is its inability to identify nonlinear correlations.

### 2.3.2 Principal Component Analysis (PCA)

PCA is a dimensionality-reduction method [88] that is often used to simplify large data sets by transforming a large set of variables into a smaller one that still contains most of
the information present in the original set. This method is particularly useful for datasets that have a large number of attributes, where reducing dimensionality can help simplify the structure of the dataset. A trade-off associated with this technique involves a loss of some of the data accuracy. Ideally, PCA can transform a dataset (with a large number of variables) into a lower dimensional one, while preserving as much of the original information as possible. Typically, a reduced dataset is preferred over a higher dimensional one due to much better convenience of visualizing or analyzing datasets with small number of variables.

The first step of the PCA procedure is to standardize the data, such that all the (continuous) variables contribute equally to the following analysis. On the next step, the covariance matrix is computed to evaluate the extent of variation of the input dataset variables away from the mean. Principal components are then constructed as new variables that are linear combinations or mixtures of the initial variables. In order to accomplish this, the eigenvectors and eigenvalues of the covariance matrix are calculated and linear combinations of the initial variables are then transformed in such a way that the new variables (principal components) are uncorrelated, while most of the information within the initial variables is squeezed or compressed into the few leading components (maximal amount of variance).

Mathematically, for example, a dataset \( X \) is represented by an \( n \times p \) data matrix whose column \( j \) is vector \( x_j \) of observations on the \( j \)-th variable. The goal of PCA is to find a linear combination of the columns of matrix \( X \) with maximum variance. The transformation is
defined by a set of size \( l \) of \( p \) dimensional vectors \((l < p)\) with weights \( \mathbf{w}_{(k)} = (w_1, \ldots, w_p) \) that maps \( \mathbf{x}_i \) in \( \mathbf{X} \) to a new vector of principal components with scores \( \mathbf{t}_i = (t_1, \ldots, t_l)_i \), where \( t_{k(i)} = \mathbf{x}_i \cdot \mathbf{w}_k \), \( i = 1, \ldots, n \) and \( k = 1, \ldots, l \). Here \( \mathbf{X}^T\mathbf{X} \) is the covariance matrix of the data and the weights represented by \( \mathbf{w} \) are the eigenvectors of the covariance matrix [22].

To maximize variance, \( \mathbf{w}_1 \) must satisfy the following criterion:

\[
\mathbf{w}_1 = \arg \max \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} / \mathbf{w}^T\mathbf{w}
\]  

(2.28)

The first principal component has a score \( t_{1i} = \mathbf{x}_i \cdot \mathbf{w}_1 \) and subsequent components can also be computed in a similar fashion.

While PCA is a popular technique that has been utilized in numerous investigations, this procedure is also limited to identifying linear correlations only. Hence, datasets with suspected non-linear relationships between variables need to be studied with more advanced data reduction techniques (such as, e.g., t-stochastic neighbor embedding [89] or multidimensional scaling [90]).

2.3.3 Median Analysis

To retrieve an overall snapshot for the dataset in question, it is important to evaluate its data properties, such as centrality and variance associated with it. For measuring central tendencies, a holistic approach, such as finding the median of each attribute or descriptor value for the entire dataset can be used. The degree to which the data tends to spread around its mean is usually measured by the standard deviation. Continuing the example
Figure 2.5: (a) Schematic of skewness of a dataset describing how much a distribution differs from a normal distribution, either to the left or to the right. The skewness value can be either positive, negative or zero and a perfectly normal distribution will have 0 skewness. (b) A dataset is divided into two sets: below median value (gray), above median value (orange) using median analysis. The difference between these two sets for a descriptor is represented by the blue bar. Larger differences observed for any particular descriptor are likely to produce an increased variance in the end point when that descriptor is used for construction of ML models.
with the university database, we can consider a dataset containing records of students registered for a particular course, with attributes including age, year in school, grades of prerequisites taken, attendance, as well as homework, daily quiz and final exam scores, plus the overall grade received, which can be viewed as the ‘target’ property of interest. For each of these attributes, after finding the median it is possible to divide the entire dataset into at least two new subsets, one containing entries >= median and the other with entries < median values. The data contained within these two subsets will include student records with corresponding attributes and once plotted, one can easily identify which attributes have contributed the most to the final grade — by computing the average final grade and standard deviation for each subset. This type of analysis helps one to find out if the data within the dataset is distributed symmetrically or skewed (biased), as well as to identify descriptors that may carry more weight in predicting the target property as shown in Fig. 2.5.

2.3.4 Traditional ML algorithms

 Regression and classification algorithms are the two major categories used for construction of predictive models in the supervised learning mode. If the target property is a continuous variable, then regression algorithms are preferred, whereas classification algorithms are useful when discrete outcomes are evaluated. For example, a prediction of a temperature variable is a good match for a regression technique, while distinguishing a
The type of fruit among apples, oranges, etc. is a generic instance of a classification problem.

### 2.3.4.1 Regression Algorithms

The simplest algorithm of this type is built on linear regression, where the response function is represented by a linear combination of independent features. The ML model
can then be optimized by using the loss function that minimizes the residual sum of squared errors (least squared errors). This type of L1 normalization is used in LASSO with an added penalty equivalent to absolute value of the magnitude of the coefficients accompanying the features. Therefore, LASSO can estimate the importance of features related to the predicted response based on the largeness (or smallness) of their coefficients. Ridge regression algorithms use L2 normalizations for optimization of models, which is also an effective way of achieving numerical stability and increasing predictive performance. This type of normalization adds a factor proportional to the sum of squares of the coefficients to the loss function. Both KRR and SVMR algorithms are similar in terms of utilizing kernels (non-linear functions) but differ in their loss functions. While KRR follows the L2 normalization, SVMR has an epsilon-insensitive (function ignoring the errors found within a certain distance of the true value) loss function. The value of epsilon defines a margin of tolerance where no penalty is given to errors. RFR is an ensemble-learning type algorithm that uses multiple decision trees run in parallel. A decision tree breaks down the data into smaller and smaller subsets by making recurrent decisions based on asking a series of questions. As shown in Fig. 2.6, it is a flowchart-like structure, where each internal node represents a “test” of an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent decision rules. In a typical RFR algorithm, a bootstrapping (sampling with replacement) technique is employed to aggregate the outputs from each
2. Computational Methods

decision tree to obtain an average prediction for the response variable.

2.3.4.2 Classification algorithms

Classification-based predictions involve a task of approximating a function $f$ mapping input variables $X$ into discrete output variables $y$ that can be also called labels or classes. One of the simplest classification algorithms is a logistic regression that estimates the probability of an occurrence of an event based on one or multiple inputs. A decision tree based classification algorithm uses an if-then rule set, which is mutually exclusive and exhaustive for classification. Here, all the attributes are categorical and the tree is constructed in a top-down recursive divide-and-conquer manner. The conditions or rules are learned sequentially using the training data. This process is iterated over the entire training set until it has reached a terminal condition.

The kernel types, value of epsilon, number of decision trees and cross-validations are the main hyperparameters that can be chosen and tuned in the process of constructing a performance-optimized ML model. Cross-validations are conducted to ensure the generated models are not over-fitting data. Cross-validation partitions a dataset into mutually exclusive subsets and uses one of them for training and others for testing. This procedure can be repeated over all reasonable subset splittings of the main dataset. For small to medium size datasets, ML models based on any of the aforementioned regression or classification approaches can be quickly constructed and trained, however, for evaluating and understanding input variables importance to the endpoint, RFR and LASSO meth-
ods are highly preferred. All ML models presented in Chapter 4 were developed utilizing algorithms discussed here, as implemented in R version 3.4.2 [91] and Python (version 2.7 or higher) [92, 93].

2.3.5 NN algorithms

NNs approximate universal functions[94] to establish relations between the input and output variables present in a dataset. The basic building block of a NN is a neuron which has the same functionality as compared to neurons in a human brain. Each neuron is a small computing unit that takes a set of real valued numbers as input, performs computations on them, and produces a single output value. Every input (x) to a neuron (also known as a perceptron) has an associated weight (w), which is assigned on the basis of its relative importance to other inputs. In the mathematical model of a neuron, if the weighted sum of inputs is greater than a specific threshold, it would give an output 1, otherwise an output 0. Every neural unit takes in a weighted sum of its inputs, with an additional bias term. Activation functions are also introduced in NNs to consider non-linearity into the output of neuron. There are several activation functions as discussed below that are commonly used in practice.

1. **Sigmoid**: It takes a real valued input and maps the output into the range [0,1]. This is similar to a sigmoid function.
2. **Tanh**: It is very similar to the sigmoid function, but it maps the output into the range [-1,1].

3. **Rectified Linear Unit (ReLU)**: Its value is equal to x when x is positive, and 0 otherwise, similar to a linear function.

A NN is composed of multiple layers, which is a collection of neurons, with connections between different layers. These layers transform data by first calculating the weighted sum of inputs and then passing it through the activation functions to capture any non-linearity present in the data. In the case of deep NNs, where there are >2 hidden layers in NN architecture, a NN makes accurate predictions by learning the weights for each of the neurons at every layer utilizing the back-propagation algorithm [95]. Back-propagation enables fine-tuning of the weights of a NN based on the error rate (i.e. loss) obtained in the previous epoch (i.e. iteration). Proper tuning of the weights ensures lower error rates, making the model reliable by increasing its generalization.

There are multiple NN algorithms[94] that are popular and we have included a brief discussion on some of these algorithms below.

**Fully-connected Feed-forward NNs** (also known as multilayer perceptrons) are constructed using multiple layers of neurons. In this type of network, all neurons of one single layer are connected to the neurons in the subsequent layer, making the network fully-connected. Training perceptrons usually requires back-propagation, giving the network paired datasets of inputs and outputs. Due to their parallel structure, the computa-
tion speed is high for this type of NNs. **Convolutional NNs (CNN)** are primarily used for image processing and are quiet different in nature as compared to the other types of NN algorithms. This algorithm takes in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The objective of the convolution operation is to extract the high-level features such as edges, from the input image. CNN algorithm uses back-propagation in a feedforward net with many hidden layers, many maps of replicated units in each layer and then pooling of the outputs of nearby replicated units to train the NN. Another type of NN is **recurrent NNs (RNNs)** where neurons retain the information of the previous iteration across time. RNNs are extremely powerful, because they combine two properties such as: (a) distributed hidden state that allows them to store a lot of information about the past efficiently; and (b) non-linear dynamics that allows them to update their hidden state in complicated ways. A popular variant of RNN is **Long Short-Term Memory (LSTM)**-based NN where an addition cell layer is introduced between the hidden states to make sure the transfer of hidden state information from one iteration to the next is reasonably high. These are very useful in modeling complex sequences or time series. **Autoencoders** are another type of unsupervised NNs designed where the data is not labeled. An Autoencoder has an encoder and decoder associated with it. An encoder compresses the data by reducing the given high-dimensional input into a low-dimensional latent representation. A decoder can then be utilized to reconstruct the input back from the encoded version. The mechanism of Autoencoders are analogous to that of PCA, the difference being
that they can use non-linear transformations to encode the given vector into smaller dimensions whereas PCA only uses linear transformations. Hence, Autoencoders can also be used in dimension reductions. A **Generative Adversarial Network (GAN)** is also a popular NN architecture in which two networks (generally a feed-forward and/or CNN) work together. There are primarily a couple of tasks of generation and discrimination of data distributions that rule how GANs work. The generator part models a transform function. It takes as input a simple random variable and must return, once trained, a random variable that follows the targeted distribution. As it is very complicated and unknown, the discriminator (discriminator function) is modeled using another NN. It takes as input a point and returns as output the probability of this point to be a “true” one. The goal of the generative NN is to maximize the final classification error between the true and generated data. The goal of the discriminator is to detect fake generated data by minimizing the final classification error. At each iteration of the training process, the weights of the generative network are updated in order to increase the classification error whereas the weights of the discriminative network are updated so that to decrease this error. There is often a competition between the two NNs which ultimately improves the performance of the GAN.

Many of these algorithms have been utilized in the domain of materials science to evaluate numerous materials property as well as predict novel materials [23–30]. Due to limited datasets size, we have only utilized traditional ML algorithms in our studies as reported in Chapter 4.
Figure 2.7: Representative plots of predicted versus observed spin moment size ($\mu_B$) produced by a ML model with 192 data points. The models in the left and right panels use the same dataset, with the only difference being its partitions into the training and test subsets. The size of the test subset is the same in both cases, however, the predicted RMSEs differ by $\sim 5\%$. $E_{Tr}$ and $E_{Ts}$ symbols refer to, respectively, average training and test subset RMSEs.

2.3.6 Learning Curves

We note that one of the main difficulties associated with ML model building, testing, and validation is the selection of appropriate subsets used on each of these steps. This problem is especially prominent for small datasets, where limited amount of available data may severely impact the ability to assess the inherent accuracy of the built model. A typical example is shown in Fig. 2.7, highlighting that standard deviations in the predicted value of the target property may be widely different for the same ML model and dataset, but different splittings of the latter into training and test subsets. Therefore, any accuracy assessments made from a single unique partitioning of the dataset into testing...
Figure 2.8: Representative plot of comparative learning curves for a ML model constructed with an RFR algorithm. $E_{\text{Tr}}$ and $E_{\text{Ts}}$ symbols refer to average training and test subset RMSEs, respectively. Note that here $E_{\text{Tr}} \ll E_{\text{Ts}}$, which is typical for a situation with a small dataset size. As the size of the dataset (and consequently the size $N$ of the available training set) increases, the $E_{\text{Ts}}(N)$ curve may approach the $E_{\text{Tr}}(N)$ curve from above.

and training sets would be suspicious. In order to avoid any statistical bias, we utilized a learning curve approach in all of our ML-based studies. This approach allows us to evaluate the performance and accuracy of the developed ML models in a consistent fashion, by visualizing the dependence of the model RMSE on the training set size.
For each point on the learning curve, such as ones shown in Fig. 2.8, the average RMSE is calculated using 1000 randomly generated (by sampling with replacement) training and test set evaluations. The average RMSE for a training set size $N$ is denoted by $E_{\text{Tr}}(N)$, whereas for the corresponding test set it is denoted by $E_{\text{Ts}}(N)$, although the size of the test set is the total number of points minus $N$. For the chosen value of $N$, the test set RMSE provides the expected error in predicting the target property of interest for the given model, while the difference between the $E_{\text{Ts}}(N)$ and $E_{\text{Tr}}(N)$ curves is an estimation of how much variance or overfitting the model contains. Following this convention, the learning curve approach has been utilized in all the projects described in Chapter 4 and the average root mean square or mean absolute errors are reported to quantify the predictive accuracy of each ML model built.

### 2.4 ML algorithms in the materials domain

The limitation on available dataset sizes is very common in ML-based investigations involving materials, which is one of the primary reasons behind the popularity of conventional ML algorithms (classification and regression) in this area of research. Furthermore, any mappings established between materials structure and properties are easier to interpret in workflows designed around these algorithms. In cases when large datasets are available, there have been some efforts in training neural networks for investigations of materials structure and properties [2, 27, 28]. However, these deep learning techniques
are commonly treated as ‘black boxes,’ where exact mathematical relationships between neurons are often very complex, which poses difficulties in understanding of physical structure-properties connections that may be suggested by such models. It is safe to assume that for accurate performance and meaningful predictive capabilities all of the supervised, semi-supervised and unsupervised learning methods [22, 94] require careful selection of features or descriptors related to physical and chemical underpinnings of the studied materials properties. Some of these descriptors — e.g., structural features of materials including atomic coordinates, bonding, crystal symmetries, etc., or their elementary electronic properties, such as band gaps — are often easily obtainable from existing experimental reports or can be computed in a quick and straightforward fashion. However, for describing systems that exhibit ‘exotic’ behavior that may involve magnetism, ferroelectricity, superconductivity and other unusual phenomena, additional complex (i.e., more information-rich) descriptors are required, whose selection, specific form and expected relationship with target properties of interest may not be apparently clear. Additionally, it would be highly advantageous if these advanced descriptors are still fairly easily obtainable — i.e., in case of computable descriptors, the associated numerical simulations should not be overwhelmingly demanding.

With these considerations in mind, we have conducted three representative ML studies, involving complex materials systems and non-trivial target properties or functionalities, such as (1) crystallization propensity of small organic molecules, (2) magnetic properties and ordering in actinides and (3) design principles for constructing new polymer-
and molecular-crystal based ferroelectrics. All of these projects are connected, in the sense that they utilize similar — and what we call *materials agnostic* — data-mining and ML approaches, although in each case the chemical space explored and target properties are quite unique. Furthermore, the developed ML-based frameworks, built on combining data of computational and experimental origins, provide new physical and chemical insights into the underlying nature of the investigated materials systems, as is further discussed in detail in Chapter 4.
Chapter 3

Human Learning

An electronic-level understanding of complex functional behavior, such as ferroelectricity or magnetism, obtained in what we call the ‘Human Learning’ mode, is still necessary as a stepping stone for exploring more elaborate data-driven learning approaches aimed at design of new materials and properties enhancement. In the following sections, we present the detailed studies of a number of different ferroelectric and magnetic compounds, conducted in the ‘Human Learning’ regime, and discuss the developed insights involving the dependence of their electronic, polar and magnetic properties on composition and structure.
3.1 Molecular- and polymer-based ferroelectric crystals

Recent advances in the synthesis of polar molecular materials have produced practical alternatives to ferroelectric ceramics opening up exciting new avenues for the incorporation of such compounds into modern electronic devices. While displaying modest polar properties, molecular- and polymer-based ferroelectric crystals provide a viable alternative by virtue of being light, flexible and environmentally friendly. In addition, a range of practical and inexpensive prescriptions is available for their synthesis and processing, as either bulk materials, or nano structures. However, in order to fully realize the potential of polar polymer and molecular crystals for modern technological applications, it is then paramount to acquire detailed molecular-level understanding of mechanisms governing the emergence of ferroelectricity in such functional materials.

3.1.1 Ferroelectricity in PVDF-TFP

This discussion is based on the manuscript titled “First-principles studies of spontaneous polarization in mixed poly(vinylidene fluoride)/2,3,3,3-tetrafluoropropene polymer crystals” by A. Ghosh, L. Louis, A. D. Asandei, and S. Nakhmanson, published in *Soft Matter* in 2018 [96].
3. Human Learning

3.1.1.1 Introduction

PVDF is the most well-known organic electroactive compound. Its popularity is attributed not only to respectable polar, piezoelectric and pyroelectric properties [97–106], but also to its structural simplicity (on the molecular, rather than microscopic level) and amenability to modification, e.g., by copolymerizing with other molecular species. Therefore, PVDF can serve as a convenient template model for exploring the nature of polymer ferroelectricity, as well as for uncovering new routes for its customization, manipulation and enhancement.

Typically synthesized as a mixture of ordered and disordered phases, PVDF requires further processing, usually in the form of stretching and poling [55, 107], to develop macroscopic polarization. Although a variety of PVDF polymorphs (some of them fictitious) can sustain polar ordering [108], the all-trans $\beta$ phase is considered to have the strongest polarization. While its magnitude is theoretically estimated to be approximately 0.18 C/m$^2$ for a fully crystalline sample [109], the crystallinity of the best experimentally grown materials is about 50–60% [110, 111], which substantially reduces their polar properties. To remediate this issue, $\beta$-PVDF is usually copolymerized with trifluoroethylene (TrFE, CHF=CF$_2$) or tetrafluoroethylene (TeFE, CF$_2$=CF$_2$), whose concentration varies from 20 to 30%. The resulting structures are more than 80% crystalline, but their polarization is reduced by around 25–30% (to 0.120–0.135 C/m$^2$), compared to the ideal crystallinity estimate above [109]. The drop in polarization happens because the copoly-
mer monomers possess either low (less than 1 D in TrFE) or nonexistent (TeFE) intrinsic dipole moments in comparison with the large dipole moment of the VDF monomer (2 D) [109].

Naturally, it would be interesting to attempt combining VDF with other fluorinated molecules, especially those that are more polar than TrFE and have compact shape, i.e., do not possess protruding branches that could hinder ferroelectric switching under applied electric fields. Coercive fields in P(VDF/TrFE) systems are very large and more than an order of magnitude higher than those in ferroelectric perovskite oxides or other classes of organic ferroelectric systems.
Here, we have employed a DFT based approach, combined with the Modern Theory of Polarization formalism utilizing MLWFs, as detailed in Chapter 2, to evaluate the polar properties of polymer crystals created by mixing $\beta$-PVDF with TFP [or HFO-1234yf commercially]. While radical homopolymerization of TFP and its copolymerization with VDF has been asserted in the patent literature more than fifty years ago [112, 113], electroactive properties of the resulting materials were never reported. A structurally similar system that has been recently reported and characterized includes a combination of VDF with hexafluoropropene (HFP), is ferroelectric at low copolymer concentrations ($P = 0.08$ C/m$^2$ at 5% HFP) and also has attractive piezoelectric properties, but somewhat lower crystallinity, compared to PVDF [114–117]. We expect that both strong spontaneous polarization and dielectric response will be observed in highly-crystalline systems based on combinations of VDF and TFP units, due to the large inherent dipole moment of the TFP monomer (2.4 D).

3.1.1.2 Structural model construction

Since the template $\beta$-PVDF crystal structure can be regarded as a pseudo-hexagonal packing of long all-trans PVDF chains, on the initial structural modeling step, we identified generic chain varieties that can be obtained by replacing some or all of the VDF units with TFP units. TFP represents a fluorinated analogue of propene, and it could likewise conceptually be synthesized in a tactic manner. While radical polymerization obviously favors atactic structures, for simplicity, as well as to avoid high computational expenses
Figure 3.2: Structural models for the (a) \( \beta \)-PVDF/\textit{iso}-PTFP, (b) \( \beta \)-PVDF/P(VDF-alt-\textit{iso}-TFP) and (c) \( \beta \)-PVDF/\textit{syndio}-PTFP crystals. Both \( xy \) and \( xz \)-plane views are shown, together with the \( a \), \( b \) and \( c \) lattice-constant designations for the orthorhombic unit cells.
associated with utilization of the DFT techniques in large supercells, we considered only highly ordered, i.e., isotactic [iso] and syndiotactic [syndio] TFP chains, as well as chains with perfectly alternating single VDF and TFP units, e.g., such as VDF-alt-iso-TFP. Some of these chain models are shown in Fig. 3.1 together with the structural representation of the template all-trans VDF chain.

For the evaluation of properties of individual chains, i.e., on the stage before they are packed into a crystal and therefore not yet subjected to any “collective polarization” effects [109], we employed chain models of finite length with the number of units varying from 1 to 6 (see Fig. 3.1). Extending the number of units beyond 6 did not change the results presented below, as by that chain length polar properties of the system, such as the average value of unit dipole moment, appeared to be saturated. In the finite-length chain models, dangling bonds at the top and bottom of the carbon backbone were terminated by hydrogen atoms.

On the next step, the developed chain structures were assembled into three-dimensional periodic arrangements resembling the idealized β-PVDF polymer crystal geometry (the latter would have \textit{Amm2} or \textit{Ama2} symmetry, depending on whether or not alternating deflections of the VDF units away from the backbone mirror plane are taken into account [118]). In the process of applying the three-dimensional periodic boundary conditions, the finite-length chains were converted into infinite chain links that are cell-periodic along the backbone direction, which is always chosen as the \textit{z} axis in this investigation. Orthorhombic unit-cell shapes were enforced for all the probed crystal structures — i.e., unit-cell
shears were forbidden during cell volume optimization — resulting in a generic \( Pm \) system symmetry, analogous to that of \( \beta \)-PVDF. In particular, three stable configurations involving combinations of \( \beta \)-PVDF and PTFP chains were identified: (a) \( \beta \)-PVDF/\textit{iso}-PTFP, (b) \( \beta \)-PVDF/P(VDF-\textit{alt}-\textit{iso}-TFP) and (c) \( \beta \)-PVDF/\textit{syndio}-PTFP. All of these models are shown in Fig. 3.2. Although, apparently, many other similar arrangements of VDF and TFP units are possible based on the original \( \beta \)-PVDF structural template, we consider these three models as representative of the common VDF and TFP unit combinations that can be used for the initial evaluation of the polar properties of such mixed systems and, consequently, their potential utility for typical applications involving these properties.

3.1.1.3 Computational techniques

\textsc{gaussian} 09 computational chemistry program suite [119] was employed to evaluate the properties of isolated finite-length polymer chains, as such systems cannot be processed with plane-wave-based DFT packages that operate under three-dimensional periodic boundary conditions. BLYP exchange-correlation functional [68, 120] was utilized for calculations of electronic energy levels, hybridized orbitals, partial atomic charges, interaction energies and molecular dipole moments. System wavefunctions were expanded into the effective-core LANL2DZ basis set developed by Dunning and coworkers [121]. For polymer crystals, we used plane-wave DFT package \textsc{quantum espresso (QE)} [122] utilizing the PBE prescription [71] for the GGA to the exchange-correlation functional and ultrasoft pseudopotentials of Vanderbilt type [75]. Energy cutoffs were set to 60 and
300 Ry, respectively, for the electronic wavefunctions and charge density. $2 \times 4 \times 4$ or $2 \times 4 \times 8$ shifted MP grids [74] were used to sample the BZs of the polymer crystals. All internal ionic positions were relaxed to force components on individual ions of less than $7.35 \times 10^{-5}$ Ry/bohr $(10^{-3}$ eV/Å). The unit-cell lattice constants were relaxed to stress-tensor components of less than 0.1 kbar while preserving the orthorhombic symmetry of the cells.

MLWFs and their charge centers were generated using the WANNIER90 code [123] and its interface with the QE package. Custom postprocessing scripts were utilized to fold the obtained MLWFs centers to their proper locations in individual VDF and TFP monomer repeat units (making sure that these units are not broken up by periodic translations of the ions) and compute their dipole moments, as well as the total polarization of the simulation cell. A 100% crystalline $\beta$-PVDF structure (total polarization of 0.18 C/m$^2$) [109, 118] was considered as the $\lambda_0$ reference system, with polarizations of the mixed crystals computed in as a change in the reference system state due to replacement of some VDF units by TFP. Some examples of the obtained MLWFs and their charge-center arrangements in VDF and TFP units, chains, and the $\beta$-PVDF/P(VDF-alt-iso-TFP) crystal are presented in Fig. 3.4.

3.1.1.4 Discussion

Isolated Chains: The average unit dipole moments in isolated $\beta$-PVDF, iso-PTFP and syndio-PTFP chains are presented in Fig. 3.3, as functions of the number of repeat units in
Figure 3.3: Dependence of the average unit-dipole moments of (a) β-PVDF, (b) iso-PTFP and (c) syndio-PTFP chains on the number of repeat units in the chain. For the VDF chain, $D_x \equiv 0$ and $|D| \equiv D_y$, therefore only the magnitude of the dipole moment is shown.

We observed quick saturation of the average unit-dipole moment magnitude with increasing number of units in the chain in all three systems. These values are 1.8 and 2.3 D for the β-PVDF and PTFP chains, respectively, with the former value being in good agreement with the results of the previous investigation (2 D) [109].

For the same chain orientation (as shown in Fig. 3.1), unlike the symmetric VDF units, individual TFP unit dipole moments have both non-zero $D_x$ and $D_y$ components. Since in the iso-PTFP chain all of the unit dipole moments are oriented in the same way, it is easy to estimate their average values, which are 0.65 and 2.2 D, respectively, for the $D_x$ and $D_y$. On the other hand, in the syndio-PTFP chain, due to alternating unit dipole moment orientations along the $x$ axis, the average $D_x$ component is small and gradually reduces

---

### Graphs

- **β-PVDF**
  - Red: $D_x$
  - Green: $D_y$
  - Purple: $|D|$

- **iso-PTFP**
  - Red: $D_x$
  - Green: $D_y$
  - Purple: $|D|$

- **syndio-PTFP**
  - Red: $D_x$
  - Green: $D_y$
  - Purple: $|D|$
to zero as the chain gets longer. The average $D_y$ component remains the same as in the iso-PTFP chain, i.e., close to 2.2 D.

For all of the chains with more than three repeat units, $D_z$ components are found to be very small, i.e., below $5 \times 10^{-3}$ D, and therefore, for the sake of clarity, they are not shown in Fig. 3.3.

**Polymer Crystals:** In all of the polymer systems, we observe no change in the $c$ lattice constant along the polymer backbone direction and an expansion of the $a$ lattice constant,
in comparison with the $\beta$-PVDF crystal. Optimized structural parameters for all of the crystals are shown in Table 3.1. Both experimentally measured [124] and computationally evaluated (with the similar setup as the one mentioned in Section 3.1.1.3) [118] lattice parameters for the $\beta$-PVDF crystal are also shown in the table for comparison. Note that the unit cells of the $\beta$-PVDF/P(VDF-alt-iso-TPF) and $\beta$-PVDF/syndio-PTFP crystals have to be doubled along the backbone direction, as shown in Fig. 3.2, compared to those of $\beta$-PVDF/iso-PTFP and $\beta$-PVDF.

In all of the polymer systems, we observe no change in the $c$ lattice constant along the polymer backbone direction and an expansion of the $a$ lattice constant, in comparison with the $\beta$-PVDF crystal. The expansion trend is modest in the $\beta$-PVDF/iso-PTFP system and more pronounced in the $\beta$-PVDF/P(VDF-alt-iso-TPF) and $\beta$-PVDF/syndio-PTFP systems. Furthermore, the latter two structures also exhibit slight expansion along the polar direction (lattice constant $b$), while, due to the uniform stacking of its iso-PTFP chains, the $\beta$-PVDF/iso-PTFP crystal displays significant contraction in the same direction. As

<table>
<thead>
<tr>
<th>System</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$-PVDF (comp) [118]</td>
<td>8.55</td>
<td>4.83</td>
<td>2.58</td>
</tr>
<tr>
<td>$\beta$-PVDF (exp) [124]</td>
<td>8.58</td>
<td>4.91</td>
<td>2.56</td>
</tr>
<tr>
<td>$\beta$-PVDF/iso-PTFP</td>
<td>9.89</td>
<td>3.62</td>
<td>2.55</td>
</tr>
<tr>
<td>$\beta$-PVDF/P(VDF-alt-iso-TPF)</td>
<td>10.83</td>
<td>5.14</td>
<td>5.06</td>
</tr>
<tr>
<td>$\beta$-PVDF/syndio-PTFP</td>
<td>11.53</td>
<td>5.14</td>
<td>5.02</td>
</tr>
</tbody>
</table>
Table 3.2: Average dipole moments (in Debye) of the VDF and TFP units in the polymer crystals. Dipole moment value for β-PVDF is taken from Ref. [109]. Positive or negative orientation of the dipole moment along the x axis is not taken into account in the averaging procedure.

| System          | Unit  | $D_x$ | $D_y$ | $|D|$ |
|-----------------|-------|-------|-------|------|
| β-PVDF          | VDF   | 0.00  | 3.00  | 3.00 |
| β-PVDF/iso-PTFP | VDF   | 0.02  | 2.67  | 2.67 |
|                 | TFP   | 1.23  | 2.51  | 2.79 |
| β-PVDF/P(VDF-alt-iso-TFP) | VDF   | 0.01  | 2.56  | 2.56 |
|                 | TFP   | 1.26  | 2.49  | 2.79 |
| β-PVDF/syndio-PTFP | VDF   | 0.00  | 2.58  | 2.58 |
|                 | TFP   | 1.22  | 2.51  | 2.79 |

A consequence, compared to the volume of the β-PVDF unit cell, the β-PVDF/iso-PTFP structure contracts in the $xy$ plane by 14%, while the β-PVDF/P(VDF-alt-iso-TFP) and β-PVDF/syndio-PTFP structures expand in the same plane by 32 and 40%, respectively.

To evaluate the polar properties of the polymer crystals, we computed average unit dipole moments of both VDF and TFP units in each crystal. The results of these calculations are presented in Table 3.2, which also includes the data for the β-PVDF crystal obtained in the previous investigation [109]. In the same study, it was shown that the dipole moment of the VDF unit changes from $\sim$2 D in an isolated chain to 3 D in the crystal and that the bulk of this enhancement is due to electronic effects (which can be detected by the movement of the MLWF centers), not the change in the ionic positions within the VDF unit. This evidence suggests that the collective electric field exerted on each unit and
'stretching' its dipole moment is much stronger in the crystal, where it is surrounded by neighbors on both the same and adjacent chains in three dimensions, compared to that of a single chain, where this unit only has neighbors from above and below.

Although in the mixed systems we do not detect the same magnitude of the collective polarization effect as in $\beta$-PVDF (from 2 D in an isolated chain to 3 D in the crystal) [109], this enhancement is still quite significant. For the VDF unit, the dipole moment magnitude is increased from 1.8 to 2.67 D in the $\beta$-PVDF/iso-PTFP structure and to $\sim$2.57 D in the $\beta$-PVDF/P(VDF-alt-iso-TP) and $\beta$-PVDF/syndio-PTFP structures. As shown in the table, in all of the mixed crystals, VDF unit dipoles remain oriented mostly along the polar $y$ direction and no substantial $D_x$ moments, e.g., due to chain rotation around the backbone direction, are observed. It is also worth pointing out that since all of the crystals considered here are eventually relaxed into the generic $Pm$ symmetry, their $\beta$-PVDF chains are allowed to and indeed do develop slight alternating dihedral tilts in the VDF unit orientations [118]. Although noticeable, these tilts remain small and do not appreciably affect the magnitude and orientation of individual VDF unit dipole moments, with their $D_x$ components being very low.

For the TFP unit, the dipole moment magnitude is increased from 2.3 to 2.79 D in all of the considered structures. Furthermore, in all cases the dipole moment orientation remains similar to that present in the TFP-based isolated chains, with the large $D_y$ component of $\sim$2.5 D. Remarkably, the $D_x$ component of the TFP unit, while remaining smaller than the $D_y$ one, is enhanced by almost 100%, compared to its value of 0.65 D determined
in an isolated iso-PTFP chain, and is close to 1.2 D. Therefore, we can confirm that, unlike the TrFE and TeFE units [109], the TFP unit remains highly polar when incorporated into the β-PVDF ‘template’ crystal structure.

In Fig. 3.5, total spontaneous polarization, as well as its separation into the $x$ and $y$ components, is presented for all of the considered polymer crystals. Spontaneous polarization of the 100% crystalline β-PVDF crystal, obtained in the previous investigation [109], is also shown for comparison. The total spontaneous polarization obtained for the β-PVDF/iso-PTFP, β-PVDF/P(VDF-alt-iso-TPF) and β-PVDF/syndio-PTFP crystals is,
respectively, 0.20, 0.18 and 0.17 C/m$^2$. I.e., in the former structure, the polarization is slightly higher than that of $\beta$-PVDF, while in the latter two structures it is about the same. The higher total polarization of the $\beta$-PVDF/iso-PTFP system relative to the other structures is most probably due to its decreased volume [see Eq. (2.20)], as discussed above.

Unlike the highly symmetric $\beta$-PVDF crystal, where the spontaneous polarization is directed strictly along the polar $y$ axis, each of the mixed crystals also develops non-zero polarization in the $x$ direction. This contribution is relatively large in the $\beta$-PVDF/iso-PTFP system due to the ordered arrangement of all of the TFP units. It becomes less pronounced in the $\beta$-PVDF/P(VDF-alt-iso-TPF) system, as the ordered $D_x$ components of the TFP units are partially canceled out by the opposing $D_x$ components of the VDF units on the same chain. Finally, the small $P_x$ observed in the $\beta$-PVDF/syndio-PTFP crystal is actually due to a minor ($\sim 10^\circ$) rotation of the syndio-PTFP chains around the backbone $z$ direction, which is allowed by the $Pm$ system symmetry and makes the $D_x$ components of the alternating TFP molecules along the chain slightly different. Note that the syndio-PTFP chain rotation is not shown in Fig. 3.2(c).

3.1.1.5 Summary

We have applied Wannier function charge center analysis within DFT to evaluate the polar properties of highly ordered polymer crystals based on $\beta$-PVDF, $\beta$-PVDF/iso-PTFP, $\beta$-PVDF/P(VDF-alt-iso-TPF) and $\beta$-PVDF/syndio-PTFP. Our investigation suggests that due to the highly polar nature of the TFP unit, all mixed systems exhibit similar or slightly
larger spontaneous polarization than perfectly crystalline $\beta$-PVDF. Therefore, if such materials were synthesized with a high degree of crystallinity, their polarization would be substantially (~30%) higher than that of typical P(VDF/TrFE) and P(VDF/TeFE) copolymer crystals, and for the $\beta$-PVDF/iso-PTFP system, comparable with perovskite-oxide ferroelectrics, e.g., BaTiO$_3$ with $P = 0.26$ C/m$^2$.

The employed methodology does not provide specific insights on the values of energy barriers for polarization switching and coercive fields, as well as the switching nature and dynamics — that are crucial for serious performance evaluation of any ferroelectric compound. However, a variety of other computational techniques working on both atomistic and more coarse-grained levels can be applied to predict the corresponding functional behavior, while advanced experimental synthesis, self-assembly and processing approaches can be utilized to control and fine tune operational properties and performance [117].

### 3.1.2 Ferroelectricity in DIPA-X

This discussion is based on the manuscript titled “Polarization canting in ferroelectric diisopropylammonium-halide molecular crystals: a computational first principles study” by L. Louis, K. C. Pitike, A. Ghosh, S. Poddar, S. Ducharme and S. Nakhmanson, published in *J. Mater. Chem.* in 2018 [125].
3. Human Learning

3.1.2.1 Introduction

In the past decade, electroactive molecular crystals have emerged as a viable alternative to both inorganic perovskite-based ferroelectric ceramics and conventional organic ferroelectrics [126], exemplified by PVDF and its copolymers with TrFE and TeFE [127–131]. However, compared to the omnipresent PVDF/copolymer family, the ‘new breeds’ of ferroelectric molecular crystals — e.g., those made up of C$_5$H$_2$O$_5$ croconic acid molecules [132] or DIPA-X complexes [133–135] — are claimed to possess substantially improved polar properties. These properties include values of spontaneous polarization $P_s$ in excess of 20 $\mu$C/cm$^2$, as well as higher ferroelectric-to-paraelectric transition temperatures $T_c$ and drastically decreased coercive fields $E_c$ — making such molecular crystals directly competitive with the best ceramic ferroelectric materials currently available.

The remarkable success of these recent discoveries [132–134, 136] accentuates the rarity of ferroelectricity within the realm of molecular compounds. Although vast numbers of non-centrosymmetric molecules do exist or could be readily synthesized, when arranged into two- or three-dimensional regular patterns they tend to orient their dipoles in anti-polar fashion, producing structures devoid of polarization [126]. Furthermore, large energy barriers associated with reorientation of molecular dipoles under the influence of applied electric field may result in extremely high coercive fields in structures that do possess spontaneous polarization, preventing ferroelectric switching [126, 137].

To gain molecular-level understanding of mechanisms governing the emergence of
ferroelectricity present in such systems, a number of hydrogen-bonded molecular-crystal ferroelectrics, e.g., those including the $\beta$-diketone enol O=C–C=C–OH moieties [136], have been thoroughly studied theoretically [138, 139] utilizing the Berry-phase method [77, 78]. However, more recent experimental and theoretical investigations of ferroelectric properties of the DIPA-halide (or DIPA-X) series of molecular crystals inspired by the original work of Fu et al. [133, 134] produced inconclusive results. Although there is some consensus in the literature about the value of $P_s$ in DIPA-Cl (9–10 $\mu$C/cm$^2$) [133, 140, 141], polarization values raging from 2–4 [142, 143] to 6–10 [141, 144], to 20+ [134, 140, 145, 146] $\mu$C/cm$^2$ have been reported for DIPA-Br, while for the DIPA-I structure claims of high (33 $\mu$C/cm$^2$) [135], low (5 $\mu$C/cm$^2$) [141] and no polarization [147] have been made (see Table 3.3 for more detailed comparison). Most remarkably, researchers employing the same computational techniques (e.g., DFT-based code VASP with GGA and Berry-phase method) obtained widely different results (6 vs. 20+ $\mu$C/cm$^2$) for the spontaneous polarization in DIPA-Br [134, 141, 145]. Therefore, it is evident that the origins of ferroelectric behavior in the DIPA-based materials have yet to be fully elucidated and dependable avenues for the enhancement and control of their useful properties still remain to be charted out.

In this work, we have utilized the MLWF formalism [79–81] to investigate the causes for the development of spontaneous polarization in the series of DIPA-X molecular crystals, with the X counter ion being F, Cl, Br or I. As described in detail below, the MLWF approach is particularly advantageous for computing polar properties in modular sys-
tems, such as polymer, oligomer and molecular crystals, and has been already used by us to analyze the behavior of ferroelectric PVDF and its copolymers [96, 109].

Our investigation produced the following insights into the molecular-level origins of spontaneous polarization in the DIPA-X crystals: (i) All of the studied DIPA-X molecular complexes, whether isolated or immersed into crystalline environment, possess large dipole moments that are in excess of 10 D. (ii) These dipole moments are arranged in an anti-polar fashion in all of the examined crystalline geometries. This arrangement is exact in non-polar structures, while in polar ones there is a small cooperative canting of the dipoles away from the anti-polar alignment, which results in non-zero spontaneous polarization. (iii) The magnitude of the developed polarization ranges from 5 to 7 \( \mu \text{C/cm}^2 \) for all of the halogen counter ions, except F, i.e., there is no pronounced dependence of the polarization value on the counter ion chemical identity. (iv) At the same time, the polarization can be strongly enhanced by manipulating the amount of the molecular-complex dipole canting away from the anti-polar alignment. Therefore, modifications of the structure and geometry of the DIPA-X systems that promote dipole canting can yield new materials with greatly improved polar properties that should be excellent candidates for a variety of advanced technological applications.

3.1.2.2 Crystal symmetry and structural models

The ‘DIPA molecule\(^{(+)}\) / halide counter ion\(^{(-)}\)’ complex (or unit), which can be characterized as a simple salt [133], is an underlying structural building block for all of the
Table 3.3: Spontaneous polarization $P_s$ (in $\mu$C/cm$^2$), room-temperature dielectric constant $\epsilon^{RT}$ and coercive field $E_c$ (in kV/cm) of some representative ferroelectric compounds arranged by the dominating polarization-emergence mechanism and compared with the properties of the DIPA-X system. Where possible, the temperature at which $P_s$ is measured is also given. All temperatures are in K. TTF-CA = tetrathiafulvalene-$p$-chloranil.

<table>
<thead>
<tr>
<th>Materials</th>
<th>Refs.</th>
<th>$P_s(T)$</th>
<th>$T_c$</th>
<th>$\epsilon^{RT}$</th>
<th>$E_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Displacive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BaTiO$_3$ (BTO)</td>
<td>[148]</td>
<td>26 (RT)</td>
<td>381</td>
<td>$5 \times 10^3$</td>
<td>10</td>
</tr>
<tr>
<td>PbTiO$_3$ (PTO)</td>
<td>[148]</td>
<td>75 (RT)</td>
<td>763</td>
<td>210</td>
<td>7</td>
</tr>
<tr>
<td><strong>H-bond ordering</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croconic acid, C$_5$H$_2$O$_5$</td>
<td>[132]</td>
<td>21 (RT)</td>
<td>400</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>KH$_2$PO$_4$ (KDP)</td>
<td>[148]</td>
<td>5</td>
<td>123</td>
<td>30</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Dipole ordering</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(VDF$<em>{0.65}$−TrFE$</em>{0.35}$)</td>
<td>[129]</td>
<td>8 (RT)</td>
<td>363</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>VDF oligomer</td>
<td>[149]</td>
<td>13 (RT)</td>
<td>6</td>
<td>1200</td>
<td></td>
</tr>
<tr>
<td>NaNO$_2$</td>
<td>[148]</td>
<td>10 (140)</td>
<td>437</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>e$^{-}$ charge transfer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTF-CA</td>
<td>[150]–[151]</td>
<td>6.3</td>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DIPA-X</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPA-Cl</td>
<td>[133]</td>
<td>8.9</td>
<td>440</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[141]</td>
<td>10.5</td>
<td>438</td>
<td>265</td>
<td>10</td>
</tr>
<tr>
<td>DIPA-Br</td>
<td>[134]</td>
<td>23</td>
<td>426</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>(thin film)</td>
<td>[143]</td>
<td>3.5 (RT)</td>
<td>419</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[146]</td>
<td>20.5 (RT)</td>
<td>425</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[141]</td>
<td>9.9</td>
<td>422</td>
<td>410</td>
<td>11</td>
</tr>
<tr>
<td>DIPA-I</td>
<td>[135]</td>
<td>33 (380)</td>
<td>415</td>
<td>$\leq 40$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[141]</td>
<td>5.2</td>
<td>368</td>
<td>110</td>
<td>12</td>
</tr>
</tbody>
</table>
DIPA-X crystals considered in this investigation. Its large dipole moment is the result of a proton transfer from H-X to diisopropylamine during the formation of the complex, which is a mechanism that is different from those presented in Table 3.3, including the electron-charge transfer in tetrathiafulvalene-\(p\)-chloranil (TTF-CA) and similar compounds [126, 150, 151] [TTF-CA exhibits a respectable spontaneous polarization of 6.3 \(\mu\)C/cm\(^2\)]. Although, on the level of unit-dipole moments packing into a regular structure, DIPA-X crystals can be at least partially classified as ‘dipole ordering’ compounds, we prefer to separate them into a category of their own, as shown in Table 3.3.

Naturally, only some arrangements of DIPA-X units and their associated dipole moments lead to an emergence of bulk polarization. In fact, most of the already synthesized DIPA-X compounds — which include X = Cl, Br and I (with the DIPA-F structure remaining virtual for now [141]) — are typically grown as non-polar crystals possessing an orthorhombic \(P2_12_12_1\) symmetry. Furthermore, all three compounds share the same general sequence of phase transitions [133–135, 140–144, 146, 147]. Upon heating, the \(P2_12_12_1\) phase is converted to a monoclinic \(P2_1\) polar phase; the DIPA-X structural-unit packings are shown in Fig. 3.6 for both phases. This transition appears to be irreversible for DIPA-Cl and DIPA-Br, while for DIPA-I the \(P2_1\) phase is either missing completely, or does not have long-time (over 24 hrs) stability and relaxes back to the \(P2_12_12_1\) phase [135, 147]. While underpinnings of the ‘as grown’ to polar transformation in the DIPA-X system clearly deserve a thorough investigation, they are not the subject of the study presented here.
Figure 3.6: DIPA-X structural-unit packings for (a) non-polar and (b) polar phases considered in this investigation [DIPA-Cl crystal is actually shown]. Note that structural unit arrangements in the $yz$-plane for both ‘as grown’ orthorhombic $P2_12_12_1$ and paraelectric monoclinic $P2_1/c$ non-polar phases look the same. The latter phase is a ‘proxy’ for the spatially disordered high-temperature $P2_1/m$ structure. For clarity, only hydrogens belonging to the ammonium ion are explicitly shown. Legend designating different atomic species is included at the bottom.

Ferroelectric switching in the $P2_1$ phase proceeds by cooperative reorientation of the ammonium cations between positive and negative directions along the $y$ axis under the influence of an applied electric field. In turn, when the $P2_1$ phase is heated above a certain critical temperature ($T_c$), it transforms into a monoclinic $P2_1/m$ non-polar phase. The emergence of the symmetry center in the latter phase has been ascribed to a disordering of the ammonium cation positions along the polar axis, which may be static or dynamic in nature [133, 135, 142]. Unlike the original $P2_12_12_1 \rightarrow P2_1$ transformation, this polar to non-polar transition is reversible, with the crystal structure returning to the $P2_1/m$ symmetry upon cooling below the $T_c$.

Although in this investigation we are mostly concerned with the properties of the polar $P2_1$ DIPA-X phase, for comparison we have utilized the same methodological frame-
work to evaluate the characteristics of the non-polar phases for some of the considered DIPA-X compounds. Initial atomic coordinates and lattice parameters for the different phases were obtained from the following sources: (i) $P2_1$ phase for DIPA-Br and $P2_1$, and $P2_12_12_1$ phases for DIPA-I were determined from diffraction experiments conducted by a complementary study [152]. (ii) $P2_12_12_1$ and $P2_1/m$ phases for DIPA-Cl and DIPA-Br were retrieved from the Cambridge Crystallographic Data Centre (CSD). Note that the $P2_1/m$ space group actually represents an aggregate symmetry of the disordered phase, where equivalent positions of the ammonium cations along the polar axis have been averaged out. Here, instead of the disordered $P2_1/m$ structure with two DIPA-X units per cell we adopted a similar $P2_1/c$ structure with four units per cell. (iii) $P2_1$ phases for DIPA-F and DIPA-Cl were derived from the DIPA-Br structure of the same symmetry by swapping the Br atoms out for the appropriate new anions and relaxing all the ionic coordinates and lattice parameters.

### 3.1.2.3 Computational details

All first-principles calculations reported below were carried out with the help of the QE package [122] utilizing the GGA to the exchange-correlation functional according to PBE prescription [71]. Ultrasoft pseudopotentials of Vanderbilt type [75] were also adopted. Energy cutoffs were set to 60 and 300 Ry, respectively, for the electronic wavefunctions and charge density. A $4 \times 4 \times 4$ shifted MP grid [74] was used to sample the BZ. Optimization of the ionic coordinates and lattice parameters was accomplished by
minimizing the ionic forces to values below 0.001 eV/Å and unit-cell residual stresses to \( \leq 0.2 \) kbar.

Since cohesion in highly polar structures, such as DIPA-X crystals, must be dominated by dipole-dipole interactions that are well described in generic DFT methodology, van der Waals (vdW) interactions were not incorporated into the DFT framework adopted in this investigation. Still, we conducted structural optimization of \( P2_1 \) DIPA-Br and DIPA-Cl systems (under the same conditions as outlined above) utilizing two different vdW-corrected functionals [153, 154]. No major structural changes expected to strongly affect the total polarization of these systems were observed, demonstrating that the GGA-PBE prescription is reasonable for describing the behavior of DIPA-X molecular crystals.

MLWFs and their charge centers were computed using the \textsc{Wannier90} code [123] and its interface with the QE package. Custom postprocessing scripts were utilized to translate MLWFs centers to their proper locations in individual DIPA-X structural units (making sure that these units are not broken up by the cell boundaries) and compute their dipole moments, as well as the total polarization of the simulation cell. The formal polarization values in all of the considered DIPA-X polymorphs were obtained after folding the coordinates of all the computed MLWF charge centers to their expected locations, i.e., close to the positions of ions, or interatomic bonds. Iterative algorithms for calculating charge center positions do not always generate converged coordinates that are automatically attached to those of ions and bonds, and instead can produce periodic images that are shifted by one or multiple \( R \) lattice vectors, with respect to the reference unit cell.
Replacing the expected charge center position $\langle r_i \rangle$ in Eq. (2.22) with its periodic image $\langle r_i \rangle + \sum_\alpha n_\alpha R_\alpha$ ($\alpha$ marks different lattice translation directions, $n_\alpha$ coefficients are integers) alters the value of $P_{el}$ by the factor of $-2e\sum_\alpha n_\alpha R_\alpha / V$, which by definition is a sum of multiple PIQs (see Section 2.1.7 for details). Therefore, choosing the simulation cell where all the DIPA-X units are intact, i.e., not broken up by the cell boundaries, and folding all of the MLWF charge centers to their expected positions within these units ensures that the corresponding formal value of $P$ includes no extra PIQs. It is well known that in certain ‘naturally non-polar’ systems formal polarization value may be non-zero [155]. Nonetheless, as shown below in Table 3.5, for all of the considered non-polar DIPA-X polymorphs, the MLWF charge center coordinate folding procedure resulted in $P \equiv 0$. The zero values of formal polarization in the non-polar DIPA-X structures were then used as reference for computing the polarization of their polar counterparts, however, no calculations were conducted for any intermediate points along the ‘polar to non-polar’ transformation path $\lambda$, since the utilization of charge center coordinate folding eliminated the need to resolve the presence of any PIQs that may have emerged during the transformation.

To contrast the polar properties of the DIPA-X molecular complexes embedded within a crystalline environment with those of isolated DIPA-X units, dipole moments of the latter were evaluated using the GAUSSIAN 09 computational chemistry program suite [119]. BLYP exchange-correlation functional [68, 120] was employed to calculate electronic energy levels, hybridized orbitals, partial atomic charges, and interaction energies. Wavefunctions were expanded using the effective-core LANL2DZ basis set developed by
Dunning and coworkers [121]. This basis provides a reasonable tradeoff between accuracy and computational expense. Furthermore, it is one of the few basis sets available in Gaussian 09 for atoms such as I and Cl.

**Dielectric properties:** Determination of the (electronic and vibrational contributions to) static dielectric permittivity tensor $\varepsilon_{\alpha\beta}^0$ requires as ingredients the knowledge of ionic Born effective-charge tensors $Z_{\tau,\alpha\beta}^*$, high-frequency dielectric permittivity tensor $\varepsilon_{\alpha\beta}^\infty$, as well as the system dynamical matrix eigenvalues and eigenvectors $\{\omega_m^2, \eta_m\}$ at the BZ-center [156–159]:

$$
\varepsilon_{\alpha\beta}^0 = \varepsilon_{\alpha\beta}^\infty + \sum_m \Delta\varepsilon_{m,\alpha\beta}.
$$

(3.1)

Here $\Delta\varepsilon_m$ is a contribution from vibrational mode $m$:

$$
\Delta\varepsilon_m \sim \Omega_{p,m}^2 / \omega_m^2,
$$

(3.2)

and

$$
\left( \Omega_{p,m} \right)_\tau^\gamma \sim \sum_{\tau,\gamma} Z_{\tau,\alpha\gamma}^* \eta_m(\tau, \gamma)
$$

(3.3)

is the mode plasma frequency. These equations remain valid for systems with structural instabilities, as long as there are no unstable (i.e., having $\omega_m^2 \leq 0$) modes at the $\Gamma$-point with substantial values of $\Omega_{p,m}$. In this work, none of the investigated polar ($P2_1$) systems possess any unstable BZ-center modes, thus allowing the static dielectric permittivity to be straightforwardly evaluated.
3. Human Learning

Table 3.4: Structural parameters for the DIPA-X systems. Experimental results produced by the complementary study [152] are also presented for some systems. $N$ is the number of DIPA-X units per cell. Crystallographic cell orientations for all of the considered phases were chosen to emphasize similarities in the mutual arrangement of the DIPA-X structural units, as shown in Figs. 3.7 and 3.8 below.

<table>
<thead>
<tr>
<th></th>
<th>Symmetry</th>
<th>$N$</th>
<th>$a$ (Å)</th>
<th>$b$ (Å)</th>
<th>$c$ (Å)</th>
<th>$\beta$ (°)</th>
<th>Volume (Å$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polar systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPA-F</td>
<td>$P2_1$</td>
<td>2</td>
<td>6.60</td>
<td>7.12</td>
<td>7.03</td>
<td>112.27</td>
<td>305.6</td>
</tr>
<tr>
<td>DIPA-Cl</td>
<td>$P2_1$</td>
<td>2</td>
<td>7.54</td>
<td>7.63</td>
<td>7.37</td>
<td>115.75</td>
<td>381.6</td>
</tr>
<tr>
<td>DIPA-Br</td>
<td>$P2_1$</td>
<td>2</td>
<td>7.86</td>
<td>8.11</td>
<td>7.90</td>
<td>116.43</td>
<td>451.1</td>
</tr>
<tr>
<td>(exp) [152]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPA-I</td>
<td>$P2_1$</td>
<td>2</td>
<td>8.50</td>
<td>8.54</td>
<td>8.42</td>
<td>118.79</td>
<td>535.7</td>
</tr>
<tr>
<td>(exp) [152]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-polar systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPA-Cl</td>
<td>$P2_1/c$</td>
<td>4</td>
<td>8.53</td>
<td>7.96</td>
<td>13.71</td>
<td>91.26</td>
<td>930.57</td>
</tr>
<tr>
<td>DIPA-Cl</td>
<td>$P2_12_12_1$</td>
<td>4</td>
<td>8.56</td>
<td>7.96</td>
<td>13.67</td>
<td>90.00</td>
<td>931.78</td>
</tr>
<tr>
<td>DIPA-Br</td>
<td>$P2_1/c$</td>
<td>4</td>
<td>8.22</td>
<td>8.59</td>
<td>13.87</td>
<td>91.03</td>
<td>978.94</td>
</tr>
<tr>
<td>DIPA-I</td>
<td>$P2_12_12_1$</td>
<td>4</td>
<td>8.49</td>
<td>9.09</td>
<td>14.09</td>
<td>90.00</td>
<td>1087.51</td>
</tr>
</tbody>
</table>

3.1.2.4 Discussion

**Crystal structure:** Lattice parameters for the polar and non-polar DIPA-X crystals considered in this investigation are assembled in Table 3.4. These parameters are in a reasonable agreement with results of other investigations, although, as is typical for GGA-PBE DFT calculations, lattice constants are overestimated slightly, when compared with those obtained in experiments. In turn, this produces larger unit-cell volumes and, consequently, somewhat lower values of total polarization [see Eqs. (2.20) and (2.22)], as discussed in detail in the next section.
For the polar $P2_1$ DIPA-Br structure, excellent agreement is obtained between the computational results presented here and the experimental measurements of the complementary study [152]. These results are in good agreement with the measurements reported by Fu et al. for $T \leq 100$ K [134], and the DFT-based simulations of Alsaad et al. [145]. For the polar $P2_1$ phase of DIPA-I, we also find a close correspondence between our computational results and those of the complementary experimental study [152]. In general, we observe a trend for expanding the unit-cell volume with increasing halogen atomic number, which is the same as the one reported by Jiang et al. (see Tables 1 and 3 in Ref. [141]).

**Ferroelectric Properties:** The calculated polar and dielectric properties for the polar $P2_1$ phases of the DIPA-X crystals are aggregated in Table 3.5. In the second column of the Table, magnitudes of individual DIPA-X unit-dipole moments are presented for the cases of the unit being embedded within the crystal and (in brackets) for the isolated unit. In the lower part of the Table, the same information is included for some of the non-polar DIPA-X crystals. The strength of the unit-dipole moment for both isolated and polar systems increases with the growing counter ion atomic number. However, regardless of being embedded in the polar or non-polar structures, unit-dipole moments are enhanced by a substantial amount, compared to the isolated case. Such enhancement presumably happens due to cooperative dipole-dipole interactions within the crystal, which is quite similar to what was previously observed in the study of $\beta$-PVDF [109].

For further analysis and visualization of the unit-dipole moments arrangement within
Table 3.5: Polar and dielectric properties of the DIPA-X molecular crystals ($P2_1$ symmetry). For the individual DIPA-X unit (column two), the absolute magnitudes of the dipole moments are shown for the ‘dipole within a crystal’ case, computed with the MLWF method implemented in QE and WANNIER90, and — in brackets — for the isolated case, computed with GAUSSIAN 09. The magnitudes of the cell dipoles (column three) represent a vector sum over the values of all the unit dipoles within the cell. The values of spontaneous polarizations $P_s = |P_s| \hat{e}_y$ in column four are extracted from those of the respective cell dipoles by dividing them by cell volumes. Spontaneous polarizations are presented together with their PIQ$_y$ ($n = 0,1,2 \ldots$). $\theta$ are the DIPA-X unit dipole canting angles out of the (101) plane. $\epsilon_{\alpha\alpha}^\infty$ and $\epsilon_{\alpha\alpha}^0$ are the diagonal components of the high-frequency and static dielectric permittivity tensors, respectively, where $\alpha = x, y, z$. Some data for the non-polar DIPA-X systems is also shown for comparison.

|                      | Symmetry | Unit | $|D|$ (D) | Cell $|D|$ (D) | $|P_s|$ ($\mu$C/cm$^2$) | $\theta$ (°) | $\epsilon_{\alpha\alpha}^\infty$ | $\epsilon_{\alpha\alpha}^0$ |
|----------------------|----------|------|----------|--------------|---------------------|------------|-------------------------------|---------------------|
| **Polar systems**    |          |      |          |              |                     |            |                                |                     |
| DIPA-F               | $P2_1$   |      | 11.34 (7.07) | 13.71 | 15.0 ± $n \times 74.6$ | 21.8 | (2.60, 2.63, 2.66) | (3.31, 5.60, 3.47) |
| DIPA-Cl              | $P2_1$   |      | 14.41 (10.37) | 7.95 | 7.0 ± $n \times 64.1$ | 14.4 | (2.44, 2.46, 2.41) | (3.38, 3.91, 3.39) |
| DIPA-Br              | $P2_1$   |      | 15.44 (11.47) | 8.46 | 6.3 ± $n \times 57.6$ | 14.1 | (2.44, 2.46, 2.41) | (3.84, 4.14, 3.38) |
| DIPA-I               | $P2_1$   |      | 16.73 (12.91) | 8.35 | 5.2 ± $n \times 51.1$ | 12.5 | (2.35, 2.43, 2.37) | (–)                 |
| **Non-polar systems**|          |      |          |              |                     |            |                                |                     |
| DIPA-Cl              | $P2_1/c$ |      | 14.60    | 0.00 | 0.0 ± $n \times 27.7$ |      |                                |                     |
| DIPA-Cl              | $P2_12_12_1$ | 14.56   | 0.01 | 0.0 ± $n \times 29.4$ |      |                                |                     |
| DIPA-Br              | $P2_1/c$ |      | 14.48    | 0.01 | 0.0 ± $n \times 26.9$ |      |                                |                     |
| DIPA-I               | $P2_12_12_1$ | 16.85   | 0.01 | 0.0 ± $n \times 25.0$ |      |                                |                     |
Figure 3.7: Unit-dipole moments arrangement in the polar $P2_1$ phase of the DIPA-X system. DIPA-Br is actually shown; dipole arrangements in the systems with other halogen counter ions look similar. Panels (a) through (c) illustrate different views of the system, while panel (d) includes a simplified layout of the dipole moments with respect to the most important structural features of the DIPA-X unit: the halogen counter ion and the ammonium group. For clarity, only hydrogens belonging to the ammonium ion are explicitly shown. Crystallographic cell outlines are depicted in all panels in grey lines. Legend designating different atomic species is included in the center.

In the DIPA-X crystals of different symmetry we have to superimpose dipole-moment vectors over all of the DIPA-X units in the simulation cell. This vector is ‘attached’ to the unit center of charge and directed from the halogen counter ion, which is the negative pole of the dipole, and towards the DIPA molecule. More specifically, we can associate the location of the positive pole inside the molecule with the N–H bond that is pointing towards the counter ion. The location of the DIPA-X unit center of charge is determined as a half-distance between its negative and positive poles, whose positions are computed
as
\[ R_{p \pm} = \left( e \sum_{\tau \in p \pm} Z_{\tau} b_{\tau} - 2e \sum_{i \in p \pm} \langle r_i \rangle \right) / \left( e \sum_{\tau \in p \pm} Z_{\tau} - 2e N_{WF}^{p \pm} \right). \] (3.4)

Here, the ‘\( p^+ \)’ set of charges is represented by all the ions and WF centers belonging to the DIPA molecule, while the ‘\( p^- \)’ set includes the halogen ion and its associated WF centers. \( N_{WF}^{p \pm} \) is the number of WFs in each appropriate set.

In Fig. 3.7 we illustrate the arrangement of DIPA-X unit dipoles in the polar \( P_{21} \) crystal. The locations of the relevant DIPA N–H bonds and the associated counter ions, forming individual dipoles, are close to, but not exactly in the (101) plane of the structure. This arrangement is generic for all of the examined polar systems with different halogen counter ions. It can be regarded as a packing of mono-molecular DIPA-X layers along the \( y \) axis, where in each individual layer all unit dipoles are pointing in the same direction. However, in two consecutive layers respective dipole moment orientations are such that their projections onto the (010) plane, i.e., their \( x \) and \( z \) components, are perfectly canceled out. What remains is a small canting of the unit-dipole moments out of the (010) plane, which is the same, e.g., always along the positive direction of the \( y \) axis, in each layer. This cooperative canting of the unit-dipole moments — represented by the canting angle \( \theta \) (see Table 3.5) — leads to an emergence of a non-zero spontaneous polarization in the \( P_{21} \) DIPA-X crystal.

For comparison, in Fig. 3.8 we present the arrangement of unit-dipole moments in the non-polar \( P_{21}/c \) and \( P_{21}2_{1}2_{1} \) systems. Same as in the polar \( P_{21} \) structure, in both
Figure 3.8: Unit-dipole moments arrangement in the (a-b) ‘as grown’ $P2_12_12_1$ and (c-d) paraelectric $P2_1/c$ non-polar phases of the DIPA-X system. Note the different structural unit $xz$-plane configurations in both phases, as presented in panels (a) and (c). DIPA-Cl is actually shown; dipole arrangements in the systems with other halogen counter ions look similar. For clarity, only hydrogens belonging to the ammonium ion are explicitly shown. Crystallographic cell outlines are depicted in all panels in grey lines. The monoclinic distortion in the $P2_1/c$ system in panel (c) is slightly exaggerated. Legend designating different atomic species is included in the center.

of these systems individual unit dipoles can be arranged into layers, with planar polarization projections canceling each other out in the two consecutive layers. However, differences between the polar and non-polar cases arise on the level of the dipole-moment orientations within each layer. In the non-polar structures, dipole moments of the two neighboring DIPA-X units within the same layer have alternating cantings along the ‘po-
lar’ axis, which results in zero spontaneous polarization. We may also speculate that similar (enhanced) values of unit-dipole moments in both polar and non-polar systems are due to cooperative dipole-dipole interactions being stronger within the same layer and weaker in between different layers.

Magnitudes of the aggregate cell-dipole moments for all of the studied polar systems are presented in the third column of Table 3.5. With the exception of the DIPA-F system, their values are smaller than those of the respective individual DIPA-X units, confirming a large extent of polarization cancellation between the units within the $P2_1$ structure, as indicated in Fig. 3.7. Furthermore, these aggregate dipole moments do not exhibit the same dependence on the halogen atomic number as the unit-dipole moments, instead remaining fixed at 8–8.5 D.

System spontaneous-polarization values, shown in the fourth column of the Table, are obtained by dividing the magnitudes of the cell-dipole moments by the respective simulation-cell volumes (see Table 3.4). Taking into account the already described trend of cell volume expansion with the increasing counter ion atomic number, the corresponding spontaneous polarizations actually slightly decrease, being approximately 7, 6.3 and 5.2 $\mu$C/cm$^2$ in DIPA-Cl, DIPA-Br and DIPA-I, respectively. These values are in excellent agreement with computational results of Jiang et al., but are around 30% lower than experimentally measured polarizations for DIPA-Cl and DIPA-Br reported in the same investigation (see Table 3 in Ref. [141]). It may be possible that the decrease in some of the computed polarization values, compared to experiments, is due to overestimation
of the cell volume by the GGA-PBE calculations. On the other hand, neither the spontaneous polarization magnitudes, nor the diminishing polarization trend vs. the increasing counter ion atomic number obtained here and in Ref. [141] agree with the other reports — both experimental and computational — claiming much larger and increasing polarizations from DIPA-Br ($\simeq 20 \mu C/cm^2$) [134, 145, 146] to DIPA-I (33 $\mu C/cm^2$) [135]. In the following section, we attempt to identify the likely origins of inconsistencies in estimation of spontaneous polarizations in these compounds.

3.1.2.5 Influence of unit dipole canting on polarization

It is tempting to resolve the mystery of divergent results of computational evaluations of $P_s$ in DIPA-Br, even when the same methodological approaches are used, by contemplating a variety of possible problems that can arise during rather delicate calculations of Berry phases. Such problems may involve poorly defined polar/non-polar state transformation paths $\lambda$ and/or the presence of unaccounted PIQs in the final results. The DIPA-X case is not the only one where the Berry-phase method may have failed to provide reasonable estimates of $P_s$ due to its inability to properly unfold the electronic phase out of the $[-\pi, \pi]$ interval it is defined upon and the MLWF technique, which can gracefully manage such issues, had to be used instead to arrive at the correct answer [160].

However, it is difficult to speculate about the validity of results of the other theoretical investigations, as the authors of Refs. [134] and [145], who obtain $|P_s| \simeq 20 \mu C/cm^2$ in DIPA-Br, do not provide sufficient information about the details of their calculations,
Figure 3.9: Dependence of polarization in the $P2_1$ DIPA-Br crystal on the hypothetical rotations of the ammonium groups, accompanied by shifting the positions of Br counter ions, represented by the change in the unit-dipole moment canting angle $\theta$. A sketch of the ammonium group rotations in two consecutive cells is included as an insert in the upper left corner of the plot.

including precisely how transformation paths $\lambda$ were constructed and how PIQs were treated. Jiang et al. at least include a statement that any PIQs have been “considered and eliminated” from their computations [141], but do not discuss why their results (which
are in nearly perfect agreement with the results presented here) are different from those obtained by the other authors [134, 145].

One noteworthy consideration is that a polar/non-polar state transformation utilizing the $P2_1$ and $P2_1/c$ structures as, respectively, $\lambda_p$ and $\lambda_{np}$ termination points has to be carried out in a simulation cell including four DIPA-X structural units (per $P2_1/c$ symmetry requirements, see also Table 3.4). The volume of such a cell would be approximately twice as large as the volume of the original $P2_1$ cell, because of the cell doubling in the (101) plane. Therefore, the value of $\text{PIQ}_y \equiv 2eb/V$ associated with the large four-unit $P2_1/c$ cell should be twice as small, compared to the one for the original two-unit $P2_1$ cell. The validity of this relationship is easy to confirm by inspecting the $\text{PIQ}_y$ values listed in Table 3.5 for polar and non-polar structures of the same halogen type, e.g., 57.6 vs. 26.9 $\mu$C/cm$^2$ for DIPA-Br. Then, in the Berry-phase polarization calculation for DIPA-Br involving the $P2_1 \rightarrow P2_1/c$ four-unit simulation cell with $\text{PIQ}_y \simeq 26–30 \mu$C/cm$^2$, polarization values of $|P_s| \simeq 6–7$ and $|\tilde{P}_s| \simeq 22–23 \mu$C/cm$^2$ fall on the same grid as $|\tilde{P}_s| \simeq |P_s| - 1 \times \text{PIQ}_y$. However, the value that is reported as formal polarization has to be the lower of the two, which for DIPA-Br appears to be $\simeq 6–7 \mu$C/cm$^2$.

3.1.2.6 Summary

We have conducted DFT computations of polar properties of the DIPA-halide series of molecular crystals, utilizing the Wannier-Function based formulation of the modern theory of polarization. In modular systems, such as the one considered here, this for-
mulation holds significant advantages over the Berry-phase methodology. Specifically, it does not require construction of polar to non-polar system transformation path (which may be unknown or even impossible to set up) and helps visualize the unfolding of the electronic phase out of the $[-\pi, \pi]$ interval. Both of these issues may produce large errors in determination of the formal value of polarization in the Berry-phase approach. Furthermore, the Wannier-Function based technique provides a way to assign dipole moments to individual structural units comprising the crystal, as was done in this investigation for DIPA-halide molecular complexes.

Our analysis of the DIPA-halide unit-dipole moment arrangements shows that these large (on the order of 10–15 D) moments are organized in an anti-polar fashion in both polar and non-polar phases of the molecular crystal. The only difference between these phases is that in the former consecutive layers of unit dipoles are canted in-phase along the same axis, while in the latter these cantings occur in anti-phase either within the same layer, or within a repeated block of neighboring layers. The unit-dipole canting angles are relatively small and the total polarizations that we obtain are 7, 6.3 and 5.2 $\mu$C/cm$^2$, respectively, for DIPA-Cl, -Br and -I. These values do not display a strong dependence on the chemical identity of the halide ion and, furthermore, for DIPA-Br and -I, they do not correspond to the recent claims of much higher polarizations. We attempt to explain this disagreement by proposing that some grown samples of these materials may be electrets, i.e., possess large quasi-permanent polarization, e.g., due to exaggerated unit-dipole moment cantings, that partially dissipates after the $P–E$ loop measurements. We
also speculate that further structural changes of the DIPA-halide system — perhaps even altering the nature of the DIPA molecule — aimed at stabilization of more pronounced dipole cantings, may lead to new families of molecular crystals with large and permanent polarization.

3.2 Derivatives of BFO

Designing new multiferroics that can overcome challenges of retaining multiferroic-ity at room temperature has attracted much interest in the condensed matter and material physics community [161]. BFO, the most popular multiferroic, undergoes two phase transitions upon cooling; one at its Curie temperature of 830°C, from a paraelectric to ferroelectric and one where it becomes magnetically ordered at its Néel temperature of 370°C. Hence, the low temperature phase with rhombohedrally distorted perovskite structure, which is both ferroelectric and anti-ferromagnetic, is considered to be the most interesting. This phase has been reported both theoretically and experimentally to show high spontaneous polarization and weak ferromagnetic ordering. The large orbital radius of the Bi 6s² lone pairs is responsible for the spontaneous electric polarization, whereas the magnetism originates from Fe 3d electrons. One of the challenges is that BFO thin films exhibit low electrical resistivity [162], which limits their application in novel multifunctional non-volatile random-access memory devices.

Recent attempts to enhance electronic and magnetic properties of BFO have shown
that the presence of common A-site dopants such as La and Sr can have a significant effect on decreasing leakage current in BFO thin films [163–167]. In addition, such modifications may also control the volatile nature of Bi atoms, whereas B-site dopants are also studied in order to enhance the magnetic ordering in BFO, which may help decrease loss and leakage [168–180]. Previous reports [181, 182] also indicate suppression of local magnetic moments and enhancement of overall magnetic ordering in presence of Sr doping, while electronic properties and size of ferroelectric polarization for La-doped BFO are comparable [183, 184] to those of pristine BFO in thin films.

On the other hand, double perovskite BFMO is a potential candidate for the highly sought single phase multiferroic system. The large orbital radius of the Bi 6s\(^2\) lone pair is responsible for BFMO exhibiting polar structural distortions and high spontaneous polarization, whereas B-site ordering of Mn and Fe contributes to its magnetic properties. With the potential to improve the ferromagnetic properties of BFO while maintaining the ferroelectricity, ordered double perovskite BFMO, in which Fe and Mn as B-site cations form a superlattice within the ABO\(_3\) perovskite structure, has attracted significant interest [185–188]. Theoretically it has been shown that the net magnetic moment in a ferrimagnetic state of a tetragonal BFMO agrees reasonably well with that measured on very thin films of BFMO [189] and the B-site Fe and Mn mixing can affect the magnetism significantly.

Furthermore, in transition-metal oxides, oxygen vacancies intrinsically exist. Although previously they have mostly been treated as structural defects, at present they are perceived as new parameters for controlling functionalities of these materials [189, 190].
Hence, effects of dopants and vacancies in BFO and its corresponding derivatives require detailed studies to address the challenges affecting real-life applications of mutiferroic perovskite oxides.

### 3.2.1 La, Sr-doped BFO


#### 3.2.1.1 Introduction

Previous studies have shown that both La and Sr can be commonly used as dopants in prototypical ferroelectrics, such as PLZT ((Pb$_{1-x}$La$_x$)(Zr$_{1-y}$Ti$_y$)O$_3$), SBT (SrBi$_2$(Ta$_2$O$_9$)), SBN (SrBi$_2$(Nb$_2$O$_9$)), BST (Ba$_x$Sr$_{1-x}$TiO$_3$) and BIT (Bi$_4$Ti$_3$O$_{12}$), to improve their characteristic functional/physical properties. In PZT thin films, La$^{3+}$ dopants help increase dielectric constant and reduce the coercive field to make it more useful for ferroelectric memory and sensing/actuating applications [192, 193]. In BIT thin films, La$^{3+}$ helps improve fatigue resistance while maintaining high remanent polarization [194, 195]. Sr$^{2+}$ in BST thin films shifts paraelectric-ferroelectric transformation temperature to near room temperature for applications in telecommunications that require highly electrically tunable dielectric response [196]. The presence of Bi$^{3+}$ improves dielectric properties of SBT.
and SBN thin films by increasing their Curie temperature [197]. Here, we focus on investigating the effects of doping, specifically with La and Sr, on electronic, magnetic and ferroelectric properties of BFO.
3. Human Learning

3.2.1.2 Computational techniques

We performed first-principles calculations using the PAW method as implemented in VASP [84, 85]. The GGA was adopted for the exchange-correlation functional, with the GGA+U method used to capture the strong correlation in $d$ localized orbitals. We have used a Hubbard parameter $U_{\text{eff}} = 2$ eV for all simulations, which was shown to be reasonable in previous calculations [198, 199]. A G-type antiferromagnetic ordering on Fe cations with spins aligned oppositely in two sublattices was considered.

We first performed the structure optimization of the rhombohedral 2 formula units of (10 atoms) BFO by relaxing the atoms steadily toward the equilibrium until the Hellman-Feynman forces are less than $10^{-3}$ eV/Å. We refer to this 2 formula unit crystal cell as the primitive unit cell hereafter. These lattice parameters were then used to construct supercells for the exploration of the dopant effects with varying concentrations. In this primitive unit cell, substitution of one La or Sr for one Bi gives rise to compound BiLaFe$_2$O$_6$ and BiSrFe$_2$O$_6$ as shown in Fig. 3.10. These two specific structures also provide us an opportunity to compare the results for 50% doped BFO as calculated in the supercell technique. For the rare-earth-based ferrites, the crystal structure can be orthorhombic [200, 201], which may originate from strong correlations among 4$f$-electrons. Therefore, we also checked the possible structure change for 50% La-doped BFO when doing the ferroelectric polarization calculations, finding that the $R3$-rhombohedral structure is still sustained (note that La has no $f$ electrons).
In the $2 \times 2 \times 2$ supercell with 16 formula units of BFO (80 atoms), we varied the concentration of dopant atoms by 6%, 25% and 50% to study the response of electronic and magnetic properties. All calculations for the primitive unit cells and supercells were performed with a $5 \times 5 \times 5$ MP [74] $k$-point mesh [198] centered at $\Gamma$ and a 500 eV plane-wave cut off energy, for which the results converged. For both 25% and 50% concentrations, we built three supercells. For each of them, the distances between the dopant atoms are varied, as shown in Fig. 3.10. For both 25% and 50% dopant concentrations, Configuration III represents the structural model where the dopants are placed farthest apart along the $[111]$ direction in the supercell. Configuration I has the dopants packed closest together, while Configuration II represents the situation in between these two extrema. Structural relaxations were performed only for pristine BFO, 6% doped, 25% and 50% doped in Configuration III, for which the spontaneous polarization exists. We chose Configuration III only because we anticipate that polarization is insensitive to the distribution of the dopants. These relaxations followed the same convergence criterion as explained previously in the section. The choice of supercells is based on experimental evidence reported elsewhere [163–180, 202–205]. These particular cases were chosen to explore the location (with respect to Fe atoms) dependence of dopant effects in these supercells. The polarizations of the La-doped supercells in Configuration III were calculated using the Berry Phase method.
3. Human Learning

Figure 3.11: Band structures of (a,d) pristine BFO, (b,e) BiLaFe\textsubscript{2}O\textsubscript{6} and (c,f) BiSrFe\textsubscript{2}O\textsubscript{6} in [-4,4] eV energy range as calculated along the high-symmetry \textit{k}-point path for a rhombohedral crystal system along (\textit{F-\Gamma-Z-L-\textit{F}}). Blue represents spin-up and red represents spin-down bands. Energy is measured with respect to the Fermi energy.

3.2.1.3 Discussion

From the band structure calculations for primitive unit cells performed along the high-symmetry \textit{k}-point path for a rhombohedral crystal system (\textit{F-\Gamma-Z-L-\textit{F}}), we find that the system remains insulating for pristine BFO and BiLaFe\textsubscript{2}O\textsubscript{6}. The indirect band gap is smaller in BiLaFe\textsubscript{2}O\textsubscript{6} as compared to BFO. The VBM is located in between \textit{F} and \textit{\Gamma} points for BFO, while the CBM is located at \textit{Z}. For BiLaFe\textsubscript{2}O\textsubscript{6} and BiSrFe\textsubscript{2}O\textsubscript{6}, the VBM and
CBM are both located at point Z. The locations of the energy levels shown in Fig. 3.11 are referenced with respect to the low-lying O 2s states, whose energy is considered to be the same in all the compounds considered in this study. This is done specifically to emphasize the upward/downward shifts of the bands in energy in the doped BFO compounds. A few electronic bands in the valence band (Γ-Z-L) region are shifted above the Fermi level for BiSrFe2O6 leading to a metallic state. This is due to the hole doping effect — when Sr atoms are substituted for Bi atoms, with Sr giving up nominally two electrons in the divalent state — which leads to the shift of the chemical potential upon minor change of the bonding nature. As shown in Fig. 3.12, this corresponds to the the Fermi energy level crossing the Fe-3d band, driving the system to be metallic. Our finding is consistent with similar electronic structure calculations for SrFeO3 in AFM state [206, 207].

For the pristine BFO, the average of the staggered magnetic moment is 4.02 \( \mu_B \) per Fe, which reduces to 3.95 \( \mu_B \) and 3.74 \( \mu_B \) for La and Sr doped BFO, respectively. Therefore, A-site dopants slightly suppress the magnetic moments in these systems. More importantly, the net magnetic moment for pristine BFO is 0 \( \mu_B \), whereas net moments for the La, Sr doped cases are 0.004 \( \mu_B \) and -0.630 \( \mu_B \), respectively.

Since calculations of spontaneous polarization demand the system to be perfectly insulating, we cannot report any polarization for Sr doped BFO. For the pristine BFO and 50% La doped BFO, the values of the polarization are 114.90 \( \mu C/cm^2 \) and 122.43 \( \mu C/cm^2 \), respectively. These are the largest reported polarizations of these systems in these specific geometric configurations. The change in polarization of \( R3c \) BiLaFe2O6 as compared to
Figure 3.12: Partial spin-resolved density (spin-up – solid and spin-down – dashed lines) of Fe-3d states in pristine BFO, BiLaFe$_2$O$_6$ and BiSrFe$_2$O$_6$. The Fermi energy level of pristine BFO is taken as zero on the energy axis.
Figure 3.13: Partial density of Fe-3d states on atoms closest to the (a) La and (b) Sr atoms for 6%, 25% and 50% (Configuration III) dopant concentrations, respectively. Total EDOS for 25% doped (d,e,f) and 50% doped (g,h,i) BFO in all three configurations. The Fermi energy level of pristine BFO is taken as zero on the energy axis.
pristine $R3c$ BFO is due to the difference in ionic radii and local charge distribution of the A-site cations.

The EDOS calculations show that introducing 6% Sr doping in supercells, which amounts to only one Sr atom in the BFO supercell, drives the system to a metallic state [208]; while the La-doped BFO remains insulating. This behavior also suggests a $p$-type electrical behavior in Sr-doped BFO, where holes are the main charge carriers. As the dopant concentration is increased from 6% to 25% and 50%, these electronic properties are more pronounced, as shown in Fig. 3.13(c-h). Even for 50% dopant concentration, La doped BFO maintains an insulating state.

For 6% dopant concentration, the Fe atom with the smallest distance 3.30Å away from the dopant has the local magnetic moment of $4.03 \mu_B$ for La and $3.98 \mu_B$ for Sr cases, respectively; while the Fe atom with distance 3.11Å away from the dopant has the local moment of $-4.03 \mu_B$ for La and $-3.96 \mu_B$ for Sr cases, respectively. In addition, the local moment on the other Fe atoms varies with their distance from the dopant site, suggesting a strong dependency on the atomic environment. The Fe atom with the farthest distance from the dopant has the largest local magnetic moment. The A-site dopants suppress magnetic moments. In our calculations, we obtained average magnetic moment of 4.04 (4.00) $\mu_B$ for 6% La (Sr) concentrations. For 25% La (Sr) concentrations, the average magnetic moments are 4.03 (3.78), 4.04 (3.88) and 4.03 (3.88) $\mu_B$ in Configurations I, II and III, respectively. For 50% La (Sr) concentrations, the average magnetic moments are 4.03 (3.75), 4.03 (3.76) and 4.03 (3.75) $\mu_B$ in Configurations I, II and III respectively. We note
that our results for 50% concentration in the Configuration III supercell calculations are consistent with those done using primitive unit cells. The change is the greatest when BFO is doped with 50% Sr and suggests that (compared to La) Sr has a greater impact on the magnetic properties. This is due to the dissimilar nature of the A-site cations in La,Sr-doped BFO as compared to pristine BFO. We also report the spontaneous polarization of pristine, 6%, 25%, 50% La-doped BFO as 115.61, 116.59, 118.71 and 120.09 µC/cm², respectively. I.e., there is a 7%-10% increase in polarization upon doping.

Overall, the electronic properties of La doped BFO in each supercell are comparable to that of pristine BFO. The differences, as shown in the comparative total EDOS plots in Fig. 3.13(a,c-h), arise due to the reduction in anisotropy of the Fe-O bond distances and local structural distortions as the dopant concentrations are increased. Moreover, for the La-doped cases, the Fe 3d-4p orbital mixing and differences in electronegativities are additional factors that enhance the spontaneous polarization and magnetic properties. Similarly, for the Sr case, the Fe ion in the FeO₆ units moves to a more symmetric position, which gives rise to a variation in electron densities. However, the electronegativity of Sr is quite low compared to that of Bi and the Fe-3d band upshift across the Fermi level makes Sr-doped BFO non-insulating.

A Bader charge analysis (shown in Fig. 3.14) was also performed to gain better insights into the electronic structure and ionic-covalent bonding character of these doped systems [209]. The normalization of volume in this technique is necessary to highlight electronic structure modifications due to exact exchange. The numbers of valence elec-
Figure 3.14: Representative Bader charge volumes of A-site cations in (a) pristine and A-site dopants in (b–d) 6%, 25%, 50% doped BFO, respectively, corresponding to Configuration III. (e) Bader charges of Fe atoms closest to the dopant atom in each case.

Only valence electrons are considered in this analysis, since the effects of core electrons on the electronic behavior [209] of these systems are expected to be minimal. Only one type of configuration (Configuration III) for 25% and 50% doping was evaluated since the overall properties do not vary significantly for different configurations. The Bader charge volumes in Fig. 3.14(a-d) are indicative of more symmetric charge distributions around the A-site dopants in BFO, as described earlier. The average Bader charge for Bi cations is $12.32e$ and $8.92$ and $8.42e$ for La and Sr cations, respectively. Therefore, the estimated valence charges of Bi, La and Sr are $2.68$, $2.08$ and $1.58$ (cf. formal valence charges of $3$, $5$ and $2$).
Figure 3.15: (a) Real and (b) imaginary parts of the dielectric function $\varepsilon(\omega)$ for the pristine and 6%, 25% and 50% La, Sr doped BFO supercells.

3 and 2$e^-$, respectively. Fig. 3.14(e) shows the Bader charges of Fe atoms closest to the dopant atoms for varying concentrations. It also implies that valence charge of Fe is close to +2 (8$e^- \sim 6e^-$) for pristine and doped BFO compounds. Charges on oxygen atoms are also reduced to $\sim -1.4$ (cf. formal valence charge of -2) for all systems to maintain charge neutrality. There is nominal difference in the Fe valence states in the pristine and La-doped BFO supercells. There is a $\sim 2%$ reduction in the local charges of Fe atoms between the pristine and Sr doped BFO, which provides an explanation for the overall reduced magnetic moment in the Sr-doped BFO case. From this analysis, we can estimate that the valence charge of the La dopant atoms is higher than that of Sr by 0.5$e$.

Finally, we present the frequency-dependent dielectric function $\varepsilon(q,\omega)$ and the refractive indices of the systems under consideration to provide insight into how La and Sr doing can improve the energy loss function $\mbox{Im}[-1/\varepsilon(q,\omega)]$ that reflects the response of
a material to an external field [210]. This computation was performed using Kramers-Kronig relation as implemented in VASP [84, 85]. From Fig. 3.15, it is clear that both the phase lag and energy loss are decreased as dopants are introduced in pristine BFO [211, 212]. The dopant sites may no longer be considered isolated at higher dopant concentrations, which may lead to a dielectric breakdown of such structures when subjected to intense electric fields. From the imaginary part of the dielectric function, the transitions of O-\( p \) electrons to unoccupied Fe-\( p \) states, O-\( p \) electrons to \( p \) states in high-energy conduction bands and other inner electron excitations can also be identified in each case. The highest value of the refractive index \( n \) for pristine BFO is reported as 2.88 at an energy of 2.22 eV. For the 6% La and Sr doped cases, these values shift to 2.82 at an energy of 1.4 eV and 2.84 at 2.12 eV, respectively. As the dopant concentration is increased to 25% and 50%, these values decrease to 2.34 and 2.67 at the same energies, respectively, suggesting the same behavior of absorption due to transition from valence to conduction bands as predicted by the frequency dependent dielectric functions.

3.2.1.4 Summary

We have found that La-doped BFO maintains an insulating state, while adding Sr drives the system to be metallic. This behavior is evident from the band structure calculations. We have also obtained an enhancement of the total polarization in La-doped BFO by 4-8% over the pristine BFO. The average magnetic moment of Fe atoms decreases as we increase dopant concentration. This behavior is also dependent on the local atomic
environment. In the structural configuration where the distances between dopant atoms with respect to Fe atoms are the smallest, the variance of the local moments is the highest. Variance in $s$ and $d$ orbitals occupancy for La and Sr and $d$ orbital occupancy for the Fe-like transition metal ion may also contribute to changes in these properties.

We note that our study provides an estimate of the magnetic moments in primitive unit cells and supercells, as compared to pristine BFO. We have not performed structural relaxations for every supercell because minimal changes in atomic positions do not affect the metallic or insulating states obtained for the doped systems. We anticipate no substantial change in the electronic properties (EDOS, band structure) if the impurity concentration is varied in La doped case. Structural relaxations were performed only for pristine BFO, 6%, 25% and 50% La doped BFO in Configuration III, for which the spontaneous polarization occurs. The interference caused by large concentration of dopants in BFO does not affect the polarization significantly. The net magnetic moment of La and Sr doped BFO is increased, compared to the pristine BFO, which improves the potential for achieving ferromagnetic-ferroelectric coupling, at least in La-doped BFO. The examined structural configurations were chosen judiciously to quantify the effects of dopants under wide changes in the surrounding atomic geometry. We reason that the use of the sample averaging technique [213, 214] for a large number of dopant atoms situated at every possible site may give us deeper insights into these effects. But such calculations become extremely expensive for relatively large impurity configurations and are therefore beyond the scope of this paper. Finally, we expect that electronic properties of other
Figure 3.16: Structural models for an 80-atom pristine BFMO supercell (a) with Bi atoms marked in purple, Fe atoms in brown, Mn is lavender and oxygen in red. B-site Fe and Mn are represented by octahedron cages. The dispersive oxygen vacant sites are shown using pyramidal cages (b) corresponding to vacancy concentrations ranging from 2.08 to 16%.

possible structural configurations with different dopant concentrations will not be distinct enough from the behavior of the systems studied here to alter the presented results significantly.

3.2.2 \textbf{Bi}_2\text{FeMnO}_6: a double perovskite

3.2.2.1 Introduction

The single-phase perovskite BFO has shown magnetic and ferroelectric orders with transition temperatures much higher than the room temperature [216]. However, its antiferromagnetic spin alignment makes the system insensitive to the applied magnetic field, which limits the application in the single phase. Sizeable magnetization has been reported in thin-film BFO. In contrast, the perovskite BiMnO$_3$ [217] exhibits ferroelectricity above the room temperature but ferromagnetism at a very low temperature. With the aim to improve the ferromagnetic properties of BFO while maintaining the ferroelectricity, ordered double perovskite BFMO in which Fe and Mn as B-site cations form a superlattice within the ABO$_3$ perovskite structure, has attracted significant research interest. Theoretically it has been shown that the net magnetic moment in a ferrimagnetic state of a tetragonal BFMO agrees reasonably well with that measured on very thin films of BFMO [218] and the B-site Fe and Mn mixing can affect the magnetism significantly [186, 187]. In transition-metal oxides, oxygen vacancies intrinsically exist. Mostly they have been considered as defects that disturb a perfect order. Nowadays they are also considered as new parameters for controlling functionalities of these materials.

In this work, we report on a first-principles study of the oxygen vacancy effect, together with the electronic correlation effect, on the magnetic and electronic properties of BFMO. Starting from an insulating BFMO, we show the evolution of magnetic moment and the change in electronic structure with the vacancy concentration. In the presence of
2.08% oxygen vacancy concentration, BFMO becomes metallic and its total magnetization decreases with increasing number of vacant sites.

### 3.2.2.2 Computational techniques

A DFT-based approach, similar to the one discussed in the previous section, has been utilized to calculate the magnetic and electronic properties of BFMO. DFT within the LSDA and the LSDA + $U_{\text{eff}}$ as implemented in VASP [84, 85] was utilized to compute the magnetic and electronic properties of BFMO. We have assumed antiferromagnetic (AFM) ordering between Fe and Mn cations for all the calculations. We used $2 \times 2 \times 1$ conventional unit-cell block of tetragonal BFMO, with each cell having 20 atoms. The lattice parameters adopted for the conventional unit cell of pristine BFMO were $a = b = 5.56\,\text{Å}$, $c = 8.06\,\text{Å}$[189]. To simulate vacancy effects in BFMO, we chose a 80-atom supercell constructed by tiling around the conventional unit cell. We introduced oxygen vacancies on various sites within the supercell, as shown in Fig. 3.16, to explore changes in electronic and magnetic properties of the system both locally and globally. The volume of the tetragonal unit cell was kept constant at 995.902 Å$^3$. Fig. 3.16 (b) shows the supercell after introducing a 16.67% concentration of oxygen vacancies, where vacant sites are marked with letter V. We first took out an oxygen atom from position marked as V12, followed by V1, V2, V11, V5, V14, V15 and V8 to study the effects of increasing vacancy concentration on the system behavior. Supercells containing 2.08, 4.17, 8.33 and 16.67% vacancy concentrations were created.
Figure 3.17: Computed band structures and spin-resolved EDOS for $U_{\text{eff}}$ of (a-b) 4 eV, (c-d) 5 eV and (e-f) 6 eV.
3.2.2.3 Discussion

The effective Hubbard parameter $U_{\text{eff}}$ was varied between 4 eV and 6 eV [188] to investigate electronic properties. The band structure calculation was performed along high-symmetry BZ $k$-point path for the tetragonal crystal system ($\gamma-X-M-\gamma-Z-R-A-M$). The system shows a transition from metallic state to insulating state at $U_{\text{eff}} = 6$ eV. The electronic states around the Fermi energy are dominated by Fe and Mn $d$ orbitals. Band structure dispersion plots are shown in Fig. 3.17 for different Hubbard parameters. Our studies suggest that $U_{\text{eff}} = 6$ eV is required for the gap opening in this system. For this value of $U_{\text{eff}}$, we observe an ‘opening’ of electronic bands in between $\Gamma$ and $Z$ points that is sufficient to make the system insulating. This gap remains indirect with the CBM located at $Z$. The insulating state is required for BFMO to exhibit ferroelectric order as desired. The total EDOS for various $U_{\text{eff}}$ shows that there is hybridization among unoccupied Mn and Fe electronic states due to correlation effects, resulting in a half-metallic structure. This behavior in tetragonal BFMO is very similar to that of monoclinic BFMO [188], where $U_{\text{eff}} = 6$ eV is also required to produce an insulating system. For $U_{\text{eff}} = 6$ eV, local magnetic moments on Fe and Mn atoms are $4.375 \, \mu_B$ and $-3.812 \, \mu_B$, respectively. The BFMO superlattice has an averaged magnetization of $8.318 \, \mu_B$ or $2.321 \times 10^{12}$ esu/cm$^3$.

Presence of oxygen vacancies has a significant effect on electronic and magnetic properties of BFMO, as shown in Fig. 3.18. The oxygen vacancies were introduced in the supercell as described above to model their influence on electronic and magnetic proper-
ties of the system. As the number of vacancies increases, the band gap closes up, driving the system into a metallic state. The local magnetic moment on Fe and Mn sublattice sites are enhanced with increase in \( U_{\text{eff}} \) from 4 to 6 eV due to \( d \)-electron correlation effects in the doped systems. Depending on the relative positions of single or multiple oxygen vacancies in the supercell, the system may exhibit a crossover or transition from VB to CB region for a fixed effective Hubbard parameter.

In order to keep the BFMO system insulating, we chose \( U_{\text{eff}} = 6 \) eV and the oxygen vacancy concentration of 2.08%. Further increase in the number of vacant oxygen sites results in closing of the band gap rendering the system metallic. The details of this insulator-to-metal transition can be obtained from examining the band structure plots computed for a number of different vacancy concentration levels of above 2.08%. The presence of the oxygen vacant sites also significantly affects magnetic properties of BFMO. The overall magnetization of the system decreases as the oxygen vacancy concentration is increased, as shown in Table 3.6.

This study provides an upper bound estimate of the critical oxygen concentration in bulk BFMO, below which the system remains insulating. We also point out that the local magnetization varies considerably depending on the relative positions of vacant sites with respect to those of the transition metal ions. For example, Mn located far away from a vacant site along the diagonal of the supercell exhibits lower magnetic moment compared to that of a Mn ion positioned near a vacant site. This behavior results in an overall decrease in magnetization of the system. It is also important to consider interferences be-
Figure 3.18: Band structure and spin-resolved EDOS plots for BFMO supercells with (a-b) 2.08, (c-d) 4.17, (e-f) 8.33 and (g-h) 16.67% oxygen vacancy concentrations.
Table 3.6: Average magnetization per BFMO formula unit under changing oxygen vacancy concentration (measured in %).

<table>
<thead>
<tr>
<th>% O vacancy</th>
<th>Total Magnetization (esu/cm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>2.847×10$^{11}$</td>
</tr>
<tr>
<td>2.08</td>
<td>2.317×10$^{11}$</td>
</tr>
<tr>
<td>4.17</td>
<td>1.675×10$^{11}$</td>
</tr>
<tr>
<td>8.33</td>
<td>1.033×10$^{11}$</td>
</tr>
<tr>
<td>16.67</td>
<td>-3.908×10$^{10}$</td>
</tr>
</tbody>
</table>

tween different vacant sites caused by large concentration of oxygen vacant sites relative to the supercell size.

We note that the use of sample averaging techniques for structural models containing large numbers of oxygen vacancies may provide more insights on the physical underpinnings of changes in the system properties. However, such calculations are usually computationally expensive, especially in the case of ionic positions relaxation. We expect that small changes in atomic positions during structural relaxations will not radically change the (metallic or insulating) state of the system with vacancies.

3.2.2.4 Summary

BFMO is a possible multiferroic material that is a good model system for investigating the influence of oxygen vacancies on its electronic and magnetic properties. We showed that a strong on-site Hubbard interaction ($U_{\text{eff}} = 6$eV) is needed for the band gap opening
in pristine BFMO. We also observed that overall system magnetization decreases with increase in oxygen vacancy concentration levels in BFMO. Furthermore, band structure calculations were performed to visualize the insulator-to-metal transition in BFMO under increasing population of vacant sites in the supercell. Variation in the \(d\)-orbital occupancy on Fe and Mn ions may be a contributing factor for the observed changes in the system magnetic properties.

We note that in this investigation we have assumed that BFMO has a long-range B-site ordering, same as in previous work on Bi\(_2\)FeCrO\(_6\) [219]. It has been reported that the synthesis of BFMO (similarly to that of Bi\(_2\)FeCrO\(_6\)) with long-range B-site ordering in either bulk or thicker film form is difficult, unlike other bulk double perovskites, such as La\(_2\)MnNiO\(_6\) [220], La\(_2\)MnCoO\(_6\) [221] and Sr\(_2\)FeMoO\(_6\) [222]. Although an authoritative study of relative stability of fully B-site ordered vs. disordered BFMO would be nontrivial, we have calculated the energy of ferrimagnetic BFMO with neighboring Fe and Mn atoms swapped on the B-sites of the supercell. The switching of one pair of Fe and Mn in the \(ab\)-plane costs an energy of about 74 meV, favoring the B-site ordering, while the switching along the \(c\)-axis saves an energy of about 55 meV.

Experimentally, a significant size of magnetic moment has been measured in BFMO ultrathin films [186, 187]. Therefore, although there is no clear experimental evidence for the long-range B-site ordering, there should still be a significant volume fraction of B-site ordering in the system, because only the Fe\(^{3+}\)-O-Mn\(^{3+}\) complex can produce magnetic moments of reasonable size. This investigation suggests that a BFMO system with B-
site ordering on the Fe and Mn sublattices and in a G-type ferrimagnetic state possesses a local magnetic moment in reasonable agreement with experimental measurements on ultrathin films. Furthermore, introducing Fe/Mn B-site mixing, which is a likely scenario for the surface of the thin film, or oxygen vacancies into the system produces a decrease in the magnetic moment. Finally, a combined effect of B-site disordering and the presence of oxygen vacancies on the BFMO system properties still remains to be explored.

3.3 Chapter Summary

The ‘Human Learning’ approach applied in this Chapter have produced some detailed molecular-level insights into mechanisms governing the emergence of spontaneous polarization in ‘soft’ functional materials. For example, while in PVDF-TFP material templates the electronegativity difference between H and F atoms plays a key role in producing polarization, the same properties in DIPA-X crystals are controlled by structural arrangements of halide and ammonium ions. From these investigations, it is quite evident that the origins for emergent ferroelectricity in molecular- and polymer-based systems are diverse — considerably more so in comparison with ‘hard’ compounds, such as ubiquitous perovskite oxides — and thus such systems constitute an interesting playground for the design and enhancement of ferroic, as well as other exotic properties. There exists a small number of theoretical and experimental reports on discovering new molecular or polymeric ferroelectrics. However, each such study is a ‘one-off’ event, while any systematic
investigations aiming to identify the most important general features of such materials that should point to a propensity for developing ferroelectricity (or multiferroicity) are yet to be conducted. In the following Section 4.3, we present our attempt to conduct such an investigation that utilizes data-driven approaches and ML algorithms applied to datasets collated from reports already available in the literature (including our own Refs. [96, 125]) to judiciously shortlist candidate structures that may possess ferroelectric functionalities from a wide space of different chemical compounds.

On the other hand, our lessons learned from the studies of BFO derivative materials utilizing DFT techniques suggest that structural parameters and presence of $d$ electrons are key for altering the electronic and magnetic properties of such systems. While magnetism in transition-metal oxides is a well explored subject, precise connections between the structure / composition and magnetic properties in other classes of materials possessing partially filled $d$ and $f$ electron subshells — e.g., such as actinide-based compounds — are currently unclear. In the following Section 4.2, we apply a combination of ML and Human Learning approaches (utilizing data from experimental literature reports and our own DFT calculations where necessary) to investigate the emergence of magnetic properties (non-collinear in nature) in actinides.

In a nutshell, the four projects reported here provide enough guidance for us to construct detailed workflows for examination of properties of complete classes or families of similar materials.
Chapter 4

Traditional ML-algorithms based learning

Depending on the availability of datasets and specifics of problems under investigation, both traditional ML and NN algorithms can be employed to learn, predict materials properties and assist in accelerating novel materials discovery. We note that the utilization of these algorithms is material-agnostic, meaning that logistically similar ML frameworks can be employed to explore very dissimilar chemical spaces. In this chapter, we discuss a number of projects, based on conventional ML methods (i.e., classification and regression), where predictive models are constructed to compute crystallization propensities of small organic molecules [Section 4.1], explore magnetic properties stemming from $f$ electrons in actinide systems [Section 4.2] and evaluate tendencies for emerging ferroelectricity in polymeric and molecular crystals [Section 4.3]. In all such studies, identification
of features important to the target properties (endpoints) and construction of appropriately optimized ML models serves as the general foundation of the developed approach, or prescription that is common for all of the investigated chemical spaces and specific materials functionalities.

4.1 Crystallization propensity of APIs


4.1.1 Introduction

The process of crystallization is widely used for the isolation and purification of specialty chemicals such as active pharmaceutical ingredients (APIs). Hence, prediction of the inherent crystallization propensities of new molecular entities is paramount for the development of more efficient approaches for their bulk synthesis. Until recently, acquisition and analyses of information related to crystallizability of low molecular weight organic molecules has been primarily accomplished by experimental screening [42, 50], including high throughput and manual methods. Such approaches have shown that crystallization success rates can be influenced by a wide variety of parameters, including those
of a structural, thermal, chemical and kinetic nature [43–49, 51]. However, the ability to predict the crystallization propensity for a wide range of organic compounds is currently lacking. This is in part due to the limited availability of crystallization data as well as information on selecting the appropriate algorithm for developing the best model. Perhaps most importantly, failed crystallization experiments are rarely reported in the literature, which creates a strong bias towards compounds with a high crystallization propensity.

Data-driven approaches utilizing ML methods could potentially serve as an invaluable tool for predicting crystallization propensities and optimizing experimental crystallization conditions of new molecular entities. A number of ML algorithms-based frameworks have already been developed to predict crystallization success rates [43–49, 51]. However, most of these models were constructed from theoretical data describing only the molecular compounds themselves, while not including potentially important variables, such as experimental conditions (temperature, solvent, seeds, degradation, counter ions, etc.). The predictive capabilities of these models has also been limited (typical values of approximately 30% RMSE are reported [43, 51] and may be insufficient for accurate in silico projections of crystallization behavior of APIs. Moreover, algorithmic factors influencing predictive accuracy have not been previously assessed.

In the present study, historical data compiled by Pfizer Inc. on the results of both successful and unsuccessful crystallization experiments (referred to as the ‘Pfizer dataset’) were utilized to assess three different machine learning algorithms for the prediction of crystallization propensities of different APIs. These algorithms include RFR, SVMR and
NN and the hyperparameters utilized to build these models are given in the Appendix (Chapter 6). In addition to the algorithms, a number of experimental factors were also considered when assessing the accuracy of different models. The reason for doing this explicitly is that often models are built using only Physiochemical properties of the API under consideration. However, as is well documented in the literature [43], external factors not solely and explicitly dependent on API properties influence crystallization rates and outcomes. One of the major external factors considered in the present study was the solvent system used. In addition, we also investigated the effects of implied supersaturation levels, presence of impurities and/or degradants, and the influence of potential seeded crystallizations. Note that the current work does not attempt to predict the polymorph outcome for a given solid form, being instead focused on predicting the crystallization propensity of a given API solely based on calculated and experimental descriptors.

The overall workflow of the project, including data curation and model construction steps, is described in Fig. 4.1. The two primary sets of models consist of those considering properties of the ‘API only’ (AO) versus those that additionally include solvent properties ‘API + solvent’ (AS). Within these model families we assessed: the different algorithms (RFR, SVMR and NN), the numbers and types of descriptors (solvent and API), and the influence of additional external experimental factors. A full list of all models constructed is given in Table 4.1.

The underlying experimental data that was used in all training and testing procedures was sourced from historical data at Pfizer Inc. The Pfizer database was divided into train-
Figure 4.1: Flowchart highlighting the main developments steps for the construction of ML models for predicting crystallization propensity.
Table 4.1: Descriptions of all machine learning models developed in this work.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Dataset Classification and Choice of Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptors Analysis</td>
<td>All 88 APIs and all chemically intuitive descriptors</td>
</tr>
<tr>
<td>AO-initial</td>
<td>78 APIs and all Physiochemical properties descriptors</td>
</tr>
<tr>
<td>AO-initial</td>
<td>78 APIs, all Physiochemical properties descriptors and all MOE-2D descriptors</td>
</tr>
<tr>
<td>AO1</td>
<td>78 APIs, all Physiochemical properties descriptors and 10 most important MOE-2D descriptors</td>
</tr>
<tr>
<td>AO2</td>
<td>Best model from above rebuilt with 66 APIs, as categorized by excluding APIs with average crystallization propensity strictly 0 and 1</td>
</tr>
<tr>
<td>AO3</td>
<td>Best AO1 and AO2 models with datasets classified under various experimental factors as mentioned below</td>
</tr>
<tr>
<td>AO3(a)</td>
<td>59 APIs after excluding single experimental factor such as: ‘no crystals found’</td>
</tr>
<tr>
<td>AO3(b)</td>
<td>64 APIs after excluding single experimental factor such as: ‘&gt;95% crystallized’</td>
</tr>
<tr>
<td>AO3(c)</td>
<td>48 APIs after excluding single experimental factor such as: ‘High propensity to crystallize due to multiple anhydrous, hydrate/solvated forms &gt;5 forms/conversion from salt to free form’</td>
</tr>
<tr>
<td>AO3(d)</td>
<td>63 APIs after excluding single experimental factor such as: ‘conversion from salt to free form’</td>
</tr>
<tr>
<td>AO3(e)</td>
<td>56 APIs after excluding single experimental factor such as: ‘very high sol in most solvents &gt;50 mg/ml’</td>
</tr>
<tr>
<td>AO3(f)</td>
<td>57 APIs after excluding single experimental factor such as: ‘color change’</td>
</tr>
<tr>
<td>AO3(g)</td>
<td>61 APIs after excluding single experimental factor such as: ‘amorphous/oil SM’</td>
</tr>
<tr>
<td>AO3(h)</td>
<td>63 APIs after excluding single experimental factor such as: ‘chemical degradation’</td>
</tr>
<tr>
<td>AO3(i)</td>
<td>60 APIs after excluding single experimental factor such as: ‘purity of SM is low’</td>
</tr>
<tr>
<td>AO3(j)</td>
<td>54 APIs after excluding both experimental factors such as: ‘chemical degradation’ and ‘amorphous/oil SM’</td>
</tr>
<tr>
<td>AO3(k)</td>
<td>54 APIs after excluding both experimental factors such as: ‘High propensity to crystallize due to multiple anhydrous, hydrate/solvated forms &gt;5 forms/conversion from salt to free form’ and ‘conversion from salt to free form’</td>
</tr>
<tr>
<td>AO3(l)</td>
<td>54 APIs after excluding both experimental factors such as: ‘chemical degradation’ and ‘conversion from salt to free form’</td>
</tr>
<tr>
<td>AS1 and AS2</td>
<td>Model AS1 has 58 API entries with &gt;=10 experimental outcomes, while model AS2 has 212 API entries with &gt;=5 outcomes</td>
</tr>
<tr>
<td>AS3 and AS4</td>
<td>Models AS3 and AS4 use same datasets as AS1 and AS2, respectively, with addition of solvent descriptors and experimental factors classification</td>
</tr>
<tr>
<td>Selected Solvents Models</td>
<td>Best AS3 and AS4 models, each rebuilt for only one of five specific solvents: methanol, ethanol, chloroform, toluene and acetonitrile. Solvent descriptors were not explicitly used for these models.</td>
</tr>
</tbody>
</table>
ing and test sets, to be used for model construction, and in some cases hold out validation sets were used for the subsequent verification of accuracy.

4.1.2 Methods

In this section, following the flowchart of Fig. 4.1, we present the details of the methodology utilized to construct and test the various types of the ML models for predicting crystallization propensity.

4.1.2.1 Data acquisition and curation

Raw data was obtained from approximately 11,000 crystallization experiments conducted between 2005 and 2017 at Pfizer Inc. with 88 organic API molecules (molecular weights ranging from 159 to 709 Da). The experimental data was curated for all compounds by: grouping experiment types, noting the solvent system employed, the nature of the starting and final materials (amorphous, crystalline or none), and also documenting additional experimental factors, such as material purity, degradation, color changes, solubility of the API in each solvent system and crystallization experiment used.

The specific crystallization experiments utilized included the following: slow and fast evaporation, heating/cooling and heating/melting, cooling crystallization, antisolvent addition, mechanical or vapor stress, slurry crystallization and multi-step/combination. The outcomes of crystallization experiments were determined by characterizing the solids produced by powder X-ray diffractometry (PXRD, primary technique) and polarized
light microscopy PLM, confirmation technique. This information was critical in assessing the crystallization score for each molecule based on experiment type and solvent system used. Since the number of experiments conducted on each API were varied as reported in the raw data, a simple ratio of positive crystallization outcomes to total trials could lead to potential bias in determining crystallization rates. This was avoided by employing a statistically based regularization method to obtain “crystallization propensity score” of between 1 (always successful) and 0 (never successful) for each set of experimental conditions.

4.1.2.2 Algorithms, descriptors and model building

The curated dataset as obtained from the original historical dataset was divided into training and test set, as used for models construction and in some cases, hold out validation sets were used for subsequent verification of models accuracy. Three different algorithms — RFR, SVMR and NN, as implemented in R [91] version 3.4.2 — were employed to build two primary sets of models consist of those considering properties of the ‘API only’ (AO) and those that additionally include solvents properties, AS. Regression was chosen over classification since the crystallization propensity scores were constructed for each molecule/system from the outset.

For the AO type models, the descriptors space included physiochemical properties, such as molecular weight, number of rotatable bonds, polar surface area, number of nitrogen and oxygen atoms, number of hydrogen bonds, difference between the number
of hydrogen bond donors and acceptors, distribution coefficient (cPFlogD) and partition
coefficient (clogP). A set of standard computational Molecular Operating Environment
(MOE) [224, 225] 2D chemical descriptors was also included in the descriptors space. We
utilized the SMILEs (simplified molecular-input line-entry system) for each molecule to
compute these descriptors. The SMILEs listed here are canonicalized, meaning they are
uniquely defined for the specific molecular structure. Descriptors relying on a three di-
mensional shape of the molecule were not used in order to avoid any potential ambiguity
to the descriptors space. On the other hand, a ‘chirality’ descriptor was constructed as
the presence (1) or absence (0) of molecular chiral centers. For the AS type models, addi-
tional experimentally determined properties of API, such as boiling and melting points,
polarizability, dielectric constant, viscosity, dipole moment, density and surface tension
were included in the descriptors space along with all other descriptors considered for the
AO models.

For both families of models, initial studies were conducted to evaluate model accuracy
against the number of descriptors used. Of the approximately 1000 initial descriptors
outlined, it was found that model accuracy did not deteriorate significantly if only the 10
most important descriptors, as assigned by the RFR algorithm, were used in each of the
model families (AO and AS).

One of the main concerns of model building, testing, and validation is the selection of
data-points used in each of these steps. For the AO-type models, the size of the Pfizer
dataset is limited to 88 APIs. This limitation may severely impact the ability to assess
the inherent accuracy of the models built. This is because the standard deviations in the predicted values for the test sets may be as large as 5–10%. Therefore any comparisons originating from a single set of predictions may be misleading. To address this issue, a learning curve approach, as described in detail in Section 2.3.6, was adopted in order to exclude any statistical bias.

4.1.3 Results and Discussion

4.1.3.1 API property trend analysis

Prior to the development of the ML models, a trend analysis was performed to determine to determine the common Physiochemical properties described above which, if any, may be significant for predicting crystallization propensity. The median value for each property was obtained for all 88 APIs, followed by division of the set into two subsets, below and above the median value. For every property, the average crystallization propensity was re-calculated on the two resultant subsets as shown in Fig. 4.2. The data in the figure is organized by a decreasing difference in the crystallization propensity of the two subsets. Therefore, properties on the left hand side should lead to a larger percentage of explained variance when used in decision-tree type algorithms, such as RFR. The properties found to have the largest effect include: molecular weight, chirality, polar surface area, number of rotatable bonds and number of hydrogen bond acceptors. In particular, APIs with molecular weight equal to or less than 418.42 show a crystallization
Figure 4.2: Average crystallization propensity for splitting the data into two sets: below the median value (blue) and above the median value (red) of the given API property with the difference given in green. The considered properties are: molecular weight (MW), chirality, number of rotatable bonds (ROTBND), number of hydrogen bond acceptors (HBACCPT), polar surface area (PSA), number of nitrogen and oxygen atoms (NOCNT), difference between the number of hydrogen bond donors and acceptors (Acptdonr), distribution coefficient (cPFlog D), partition coefficient (clogP) and number of hydrogen bond donors (HBDONR).
propensity of 82%, as opposed to 53% for APIs with a higher molecular weight. Also, achiral APIs show an average crystallization propensity of 84%, compared to that of 56% for the chiral ones.

4.1.3.2 API only models

The sequential development of the AO family of models is presented in Fig. 4.3. The AO1 model was constructed using all of the APIs and all of the descriptors to provide a baseline accuracy with minimal additional data curation. The next two models, AO2 and AO3, utilized only the 10 most important descriptors, as identified by the RFR al-
algorithm, in conjunction with the API properties described in Fig. 4.2 and also take into consideration further data curation steps as described below.

As noted by Tropsha and co-workers [226–228], the accuracy of models can suffer when data is incompatible with the model endpoint. Therefore, for the AO2 and AO3 models, we considered potential experimental biases incompatible with crystallization propensities. The most obvious bias is that solvent has a direct influence on crystallization for a given API and this cannot be captured by an API only model. This will be explored in more detail in the next section. Instead for the AO2 model we attempted to correct the dataset for insufficient supersaturation levels and potential seeded crystallizations. In this scenario the experiments would either always fail or always be successful despite the properties of the API. Therefore, we removed all APIs that have observed crystallization propensities of 0 and 1.

While this procedure may remove genuine crystallization data, it should also remove a large percentage of crystallizations that are either enhanced by seeding or suppressed by having no driving force, i.e., a system without supersaturation. In principal, the presence of this dataset would result in large activity cliffs and the model accuracy should improve if they are excluded. The same logic has also gone into building the AO3 model. In this case, it is well understood that the presence of impurities and/or degradants can inhibit or eliminate nucleation completely [229]. Therefore, model accuracy may be affected if crystallizations are carried out in their presence. In this case the historical data describes the (low) purity of the ingoing material, as well as the presence of degradants and any
observed color changes, which would further indicate the presence of degradation. If any of these conditions were noted in the results, then these APIs were removed from the AO3 model along with those already excluded from the AO2 model.

Learning curves summarizing these models are given in Fig. 4.4. The curves for the AO1 model, constructed with RFR, SVMR and NN algorithms, are shown in panel (a), whereas those for the AO3 model are given in panel (b). For the AO1 model, all of the algorithms show poor predictive accuracy (i.e., high model bias), with an approximate 40% test RMSE regardless of the training test size, which is in line with previous results [43, 44, 51]. The SVMR and NN exhibit little or no data overfitting, i.e., there is no difference between the training error and the test error. The RFR algorithm, on the other hand, shows a large amount of overfitting (∼25% RMSE for training set sizes of 20–70), as is common for this method. This suggests that adding more training set data points may improve the accuracy of RFR-based models but not the SVMR- and NN-based models, whose accuracy remained constant with respect to increasing training test sizes. Similar trends for the various ML algorithms are observed for the AO3 model. However, in this case bias is substantially improved for all of the algorithms from a test RMSE of approximately 40% to below 30%. In addition, the amount of overfitting by the RFR algorithm is also reduced to ∼12% RMSE at the maximum considered training set size. Using an extrapolation of the learning curves in the AO3 model, we estimate that the training set size needs to be approximately 1100–1200 in order to remove all overfitting.

The accuracy improvement by 10% RMSE exhibited by the AO3 model over the AO1
Figure 4.4: Comparative learning curves for (a) AO1 and (b) AO3 model variants constructed utilizing all three ML algorithms. $E_{Tr}$ and $E_{Ts}$ symbols refer, respectively, to average training and test set RMSEs.
model indicates that the presence of impurities and/or degradants does impact the ability of AO-type models to predict crystallization propensities. Therefore care should be exercised in model development or data collection where such situations could occur. The AO2 model (data not shown) did not exhibit any improved accuracy over that of the AO1 model.

A detailed list of the 10 Physiochemical and 10 MOE-2D descriptors that are the most important for predicting crystallization propensity by the best of AO models, AO3, as identified by the RFR algorithm, is provided in Table 4.2.

The AO3 model identifies chirality and molecular weight as the two most important descriptors, which correlates well with the analysis presented in Fig. 4.2. The tendency of chiral molecules to have lower crystallization propensity could be justified by the reduced number of potential space groups available for these molecules to crystallize in (65 instead of 230). In a similar fashion, larger and more flexible molecules as represented through their molecular weight and number of rotatable bonds (5th ranked descriptor) have been associated previously with difficulty to crystallize. This phenomenon is related to an increased free energy barrier to crystallization that may be due to the molecule’s ability to adopt a large number of conformational sub-states. For the other AO models, although rankings of individual physiochemical descriptors may vary slightly, these descriptors and others, such as number of hydrogen bond donors and acceptors, and the partition coefficient are always identified as most important ones related to predicting crystallization propensity.
4. Traditional ML-algorithms based learning

Table 4.2: List of 20 most important API descriptors selected by the RFR algorithm for the best AO model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Importance</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chirality</td>
<td>Physiochemical</td>
<td>0.512</td>
<td>Presence of chiral centers</td>
</tr>
<tr>
<td>MW</td>
<td>Physiochemical</td>
<td>0.410</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>cPFlogD</td>
<td>Physiochemical</td>
<td>0.144</td>
<td>logD</td>
</tr>
<tr>
<td>Acptdonr</td>
<td>Physiochemical</td>
<td>0.141</td>
<td>Difference between number of Hydrogen Bond Donors and Acceptors</td>
</tr>
<tr>
<td>ROTBND</td>
<td>Physiochemical</td>
<td>0.138</td>
<td>Number of Rotatable Bonds</td>
</tr>
<tr>
<td>PSA</td>
<td>Physiochemical</td>
<td>0.137</td>
<td>Polar Surface Area</td>
</tr>
<tr>
<td>HBACCPT</td>
<td>Physiochemical</td>
<td>0.130</td>
<td>Number of hydrogen bond acceptors</td>
</tr>
<tr>
<td>PEOE_VSA_FPOL</td>
<td>MOE 2D</td>
<td>0.129</td>
<td>Fractional polar van der Waals surface area</td>
</tr>
<tr>
<td>PEOE_VSA_FPOS</td>
<td>MOE 2D</td>
<td>0.126</td>
<td>Fractional polar positive van der Waals surface area</td>
</tr>
<tr>
<td>PEOE_VSA.0.2</td>
<td>MOE 2D</td>
<td>0.123</td>
<td>Direct Electrostatic Interactions</td>
</tr>
<tr>
<td>Gao_4cav</td>
<td>MOE 2D</td>
<td>0.110</td>
<td>Kier &amp; Hall Connectivity</td>
</tr>
<tr>
<td>SMR_VSA3</td>
<td>MOE 2D</td>
<td>0.107</td>
<td>Partial Charge</td>
</tr>
<tr>
<td>kS_ssCH2</td>
<td>MOE 2D</td>
<td>0.104</td>
<td>Sum of E-state values</td>
</tr>
<tr>
<td>Gao_4pcv</td>
<td>MOE 2D</td>
<td>0.098</td>
<td>Kier &amp; Hall Connectivity</td>
</tr>
<tr>
<td>Rings*</td>
<td>MOE 2D</td>
<td>0.095</td>
<td>Topological</td>
</tr>
<tr>
<td>SlogP_VSA4</td>
<td>MOE 2D</td>
<td>0.089</td>
<td>Subdivided surface areas</td>
</tr>
<tr>
<td>PEOE_VSA_PPOS</td>
<td>MOE 2D</td>
<td>0.086</td>
<td>Total polar positive van der Waals surface area</td>
</tr>
<tr>
<td>HBDOANR</td>
<td>Physiochemical</td>
<td>0.082</td>
<td>Number of hydrogen bond donors</td>
</tr>
<tr>
<td>NOCNT</td>
<td>Physiochemical</td>
<td>0.081</td>
<td>Number of Nitrogen and Oxygen atoms</td>
</tr>
<tr>
<td>ClogP</td>
<td>Physiochemical</td>
<td>0.078</td>
<td>Measurement of Partition Coefficient</td>
</tr>
</tbody>
</table>

Finally, the predictive capabilities of the AO3 model were further assessed on the hold-out validation set containing 10 APIs not used during the model development. The average RMSE of these predictions $E_V$ is estimated as $\sim 28\%$ in line with the test set error. We note that these predictions were made in a similar fashion to that used in the learning curve approach, i.e., the average error from 1000 model predictions using sampling with replacement.
Figure 4.5: Flowchart showing the development steps for the series of the AS models.
4.1.3.3 API + solvent models

As detailed in the previous section, the best AO model can predict API crystallization propensity with $\sim 30\%$ test or validation RMSE, which is similar to the results obtained in previous investigations [43, 44, 51]. Importantly, the development of the AO models demonstrated that certain experimental factors, when explicitly accounted for, can increase model accuracy. In particular when considering the presence of impurities and/or degradants, the resultant accuracy improved by up to an additional RMSE of 10%. Another critically important experimental factor is the solvent, which has a well-known impact on crystallization kinetics. Therefore, in this section the models were built to explicitly account for this variable. The workflow employed for model building and testing of the AS models is given in Fig. 4.5. As with the AO models, we adopt the learning curve approach to access the accuracy of the various procedures. However, in this case the model end-point has become the crystallization propensity for a specific combination of an API/solvent pair. A complete list of solvents considered is included in Table 4.3.

The simplest AS that can be built use the same API descriptors as outlined in the previous section with the addition of a solvent ID descriptor. Models AS1 and AS2 use this approach, with the difference between them being the number of experimental repeats required for data inclusion. In the case of AS1 we use 10 or more repeats, whereas for the AS2 model we only require 5 repeats.

Models AS3 and AS4 are equivalent to the AS1 and AS2 models in terms of the exper-
Table 4.3: List of single solvents with their corresponding IDs used in the construction of the ‘API + solvent’ family of models.

<table>
<thead>
<tr>
<th>Solvent ID</th>
<th>Description</th>
<th>Solvent ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-Butanol</td>
<td>21</td>
<td>Diethyl ether</td>
</tr>
<tr>
<td>2</td>
<td>1-Methyl pyrrolidone</td>
<td>22</td>
<td>Diisopropyl ether</td>
</tr>
<tr>
<td>3</td>
<td>1-Propanol</td>
<td>23</td>
<td>Dimethyl sulfoxide</td>
</tr>
<tr>
<td>4</td>
<td>1,1,1-Trichloroethane</td>
<td>24</td>
<td>Dimethylacetamide</td>
</tr>
<tr>
<td>5</td>
<td>1,2-Dichloroethane</td>
<td>25</td>
<td>Dimethylformamide</td>
</tr>
<tr>
<td>6</td>
<td>1,4-Dioxane</td>
<td>26</td>
<td>Ethanol</td>
</tr>
<tr>
<td>7</td>
<td>2-Butanol</td>
<td>27</td>
<td>Ethyl acetate</td>
</tr>
<tr>
<td>8</td>
<td>2-Methyl tetrahydrofuran</td>
<td>28</td>
<td>Formamide</td>
</tr>
<tr>
<td>9</td>
<td>2-Methyl-1-butanol</td>
<td>29</td>
<td>Heptane</td>
</tr>
<tr>
<td>10</td>
<td>2-Propanol</td>
<td>30</td>
<td>Hexafluoro isopropanol</td>
</tr>
<tr>
<td>11</td>
<td>2-Propyl ether</td>
<td>31</td>
<td>Hexane</td>
</tr>
<tr>
<td>12</td>
<td>2,2,2-Trifluoroethanol</td>
<td>32</td>
<td>Isopropyl acetate</td>
</tr>
<tr>
<td>13</td>
<td>Acetic acid</td>
<td>33</td>
<td>Methanol</td>
</tr>
<tr>
<td>14</td>
<td>Acetone</td>
<td>34</td>
<td>Methyl ethyl ketone</td>
</tr>
<tr>
<td>15</td>
<td>Acetonitrile</td>
<td>35</td>
<td>Methyl isobutyl ketone</td>
</tr>
<tr>
<td>16</td>
<td>Benzyl alcohol</td>
<td>36</td>
<td>Methyl tert-butyl ether</td>
</tr>
<tr>
<td>17</td>
<td>Butyl acetate</td>
<td>37</td>
<td>Nitromethane</td>
</tr>
<tr>
<td>18</td>
<td>Chloroform</td>
<td>38</td>
<td>Pyridine</td>
</tr>
<tr>
<td>19</td>
<td>Cyclohexane</td>
<td>39</td>
<td>Tert-butanol</td>
</tr>
<tr>
<td>20</td>
<td>Tetrahydrofuran</td>
<td>41</td>
<td>Toluene</td>
</tr>
<tr>
<td>42</td>
<td>Water</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.6: Comparative learning curves for (a) AS2 and (b) AS4 models constructed utilizing the RFR, SVMR and NN ML algorithms. Note the significant increase in the training test sizes that now include multiple entries with different solvents for each API, as well as much lower degrees of data overfitting for RFR in both models.
Table 4.4: Test set errors $E_{Ts}$ and differences between the training and test set errors representing data overfitting for the training test size of 50 and the largest (max) training test size, as obtained from learning curves of AS4 and AO3 for the three ML algorithms.

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>$E_{Ts}$ (max)</th>
<th>$E_{Ts} - E_{Tr}$ (max)</th>
<th>$E_{Ts}$ (50)</th>
<th>$E_{Ts} - E_{Tr}$ (50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO3</td>
<td>RFR</td>
<td>0.26</td>
<td>0.14</td>
<td>0.26</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.30</td>
<td>0.01</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.29</td>
<td>0.02</td>
<td>0.29</td>
<td>0.01</td>
</tr>
<tr>
<td>AS4</td>
<td>RFR</td>
<td>0.20</td>
<td>0.06</td>
<td>0.26</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.42</td>
<td>0.00</td>
<td>0.44</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.29</td>
<td>0.01</td>
<td>0.39</td>
<td>0.02</td>
</tr>
</tbody>
</table>

experimental repeat requirements but use explicit solvent-based descriptors instead of solvent IDs. The solvent-based descriptors include the Physiochemical properties described previously (calculated for the solvent) along with the following experimentally determined properties: boiling and melting points, polarizability, dielectric constant, viscosity, dipole moment, density and surface tension.

For the models requiring at least 10 experimental repeats per solvent (AS1 and AS3), this resulted in only 57 unique API/solvent data points, indicating that for the remaining 30 API molecules (out of the total of 88) less than 10 experiments per solvent system were carried out. Using a smaller cut-off of 5 experiments results in a significant increase to 212 data points. In order to preserve the maximum amount of the available data for construction and testing, internal (Pfizer) validation sets were not used for the AS models. Furthermore, a variant of the AS4 model was constructed, utilizing the same data selection conditions as for the AO3 model (i.e., eliminating APIs with either a 0% or 100% crystal-
lization success rate and accounting for the presence of impurities and/or degradants), which resulted in 164 API/solvent data points. However, unlike in the case of the AO models, only a marginal improvement (<2%) in predictive accuracy was observed, compared with the performance of the generic AS4 model.

The primary results of the AS model development are given as the learning curves for the AS2 and AS4 models presented in Fig. 4.6. The curves show some interesting trends. In both cases, as for the AO3 model, the RFR algorithm outperforms the other two algorithms significantly. In this case the RFR algorithm has a test set error of 20% and overfits the data by approximately 6% for both the AS2 and AS4 models. The NN and SVMR algorithms on the other hand show minimal overfitting but the test set errors are 34% and 43% in the AS2 model and 29% and 42% in the AS4 model for the NN and SVMR algorithms, respectively. Interestingly the NN algorithm for the AS4 model shows a definite downward trend as the size of the training set is increased. In all previous cases the RMSE plateaus early for this algorithm. Significantly, when compared with the AO3 model only, the RFR algorithm utilizing solvent information outperforms the best API only model; whereas the NN and SVMR algorithms have approximately equivalent or poorer performance despite an increase in training set sizes. When comparing the performance of the AS1 and AS3 models, which required 10 experimental repeats, versus the AS2 and AS4 models, which only required 5 experimental repeats very little difference in performance was observed. For example, the RFR produced curve for the AS1 model has a test set RMSE of approximately 18% in the plateau region, whereas for the AS2
As for the AO models, the most important descriptors include the API descriptors: chirality, number of rotatable bonds, polar surface area, hydrogen-bond acceptor count and molecular weight. Interestingly, only two solvent descriptors (solvent molecular weight and melting point) appear in the top ten. However, the API variables chirality and number of rotatable bonds seem to be significantly more important than those for the solvent. This may help to explain why the AS2 model essentially has the same performance when using RFR as the AS4 model, as shown in the learning curves of Fig. 4.6. In the case of the AS2 model, the only solvent descriptor used is the solvent ID, which indicates that introducing additional complexity into the model by considering solvent Physiochemical properties has little effect on the accuracy.

To explore this observation in more detail we have summarized the performance differences between the three algorithms for the best API-only model (AO3) against the best AS4 in Table 4.4. The table details the test set RMSE for the three algorithms at a test set size of 50 and also at the maximal test set size. The table shows that at the maximal test set size, the AS4 model is superior both in terms of total RMSE and overfitting RMSE when using RFR. The other two algorithms showed a decrease in their predictive performance when utilizing solvent information (as described previously). However, when comparing relative performances of the two models (AO3 and AS4) at the same test set size, the performance difference when using RFR disappears and the difference between AO3 and AS4 is not apparent. This suggests that the predominant factor improving the
4. Traditional ML-algorithms based learning

AS models is the size of the training data and not necessarily introducing a more detailed picture of the physical aspects of the process. This was further highlighted by the results produced by the AS1 and AS3 models, where the learning curves demonstrated similar accuracy to that of the AS2 and AS4 models at the same test set size. This emphasizes the importance of understanding how ML algorithms work and that performance gains can be non-intuitive.

4.1.4 Conclusions

A regression algorithm training based ML approach has been developed for constructing models that predict the crystallization propensity of pharmaceutical compounds. Descriptors such as API molecular weight, chirality, rotatable bond counts and the number of hydrogen bond acceptors are the most important parameters for predicting crystallization propensities.

Comparative analysis of the learning curves for different regression algorithms used in the model training demonstrates that, compared to the RFR algorithm, the SVMR and NN algorithms display low overfitting and fast convergence of the model predictive accuracy with respect to the training test size. However, the RFR algorithm ultimately provides better accuracy of prediction than the other two training techniques and, furthermore, it is likely to perform better with larger training set sizes.

The inclusion of a more detailed physical picture into the models proved to give mixed
results. For the AO models, important gains in performance were realized when chemical degradation and impurities were excluded from the training data, but for the AS models this additional step had little effect. In addition, while the AS models gave the appearance of being superior to the API only models, a more careful analysis showed that this was attributed to the increased size of the training set rather than inclusion of important physical variables into the model.

4.2 Magnetism in Actinide Systems


4.2.1 Introduction

In addition to more generic traits originating from their general chemistry and radioactive behavior, lanthanide and actinide-based materials display exotic properties that originate from the presence of itinerant or localized $f$-electrons and are associated with the filling of the $4f$ and $5f$ electron subshells. In particular, the interplay of the hybridization of $5f$ electrons with itinerant conduction electrons and the on-site Coulomb repulsions among those electrons is responsible for the behavior exhibited by actinides. Such properties may include an emergence of magnetism [19, 231–233] and colossal magne-
toresistance at partial subshell fillings, as well as unconventional superconductivity and magnetism [21], metal-insulator transitions [234, 235], hidden magnetic order [236] and the presence of heavy fermions [237]. Due to strong correlation effects involving 5\textit{f}-electrons and their interactions with itinerant conduction electrons, magnetic behavior of actinide-based systems is sensitive to applied pressure, magnetic field and chemical doping. As a result, actinide-based magnetic materials are not only useful in nuclear applications but also constitute an interesting playground to push our fundamental understanding of correlated materials to the limit. So far, only the 4\textit{f}-electron magnetism has been studied with DFT-based ML tools in a general context of ternary oxide compound [238, 239]. When compared with results produced within the same computational DFT-based framework, these models can predict magnetic moment size of such systems with \(\sim20\%\) of root mean square error. However, no experimental validations for these predictions have been conducted. Furthermore, precise connections between the structure and magnetic properties in actinide-based materials are currently unknown.

In this investigation, we attempt to find such structure-property links by assembling and mining two datasets that aggregate, respectively, the results of high-throughput DFT simulations and experimental measurements for the families of uranium- and neptunium-based binary compounds. Therefore, this work involves a systematic investigation of possible connections between the structure and magnetic properties for a variety of different actinide-based binary compounds in an attempt to establish a general prescription for constructing families of ML models that incorporate computational and experimental
knowledge. Complementary utilization of data originating from both of these sources is necessary for accurate assessment and prediction of the magnetic properties of interest: average cation moment sizes (but not their ordering, which is intricately complex for theoretical simulations) can be easily extracted from DFT calculations, while magnetic ordering can be straightforwardly characterized by experiments. The twelve compounds chosen for compiling Dataset I (computational) exhibit either A-type or C-type antiferromagnetism (AFM), or ferromagnetism (FM). As further explained in Section 4.2.2.1 and validated in Section 4.2.3, we use only AFM configurations for building this dataset. Dataset II (experimental) is assembled by extracting information on magnetic ordering in uranium-based binary compounds from numerous literature sources. The acquired information serves not only as a base for constructing ML models capable of predicting magnetic ordering, but in some instances can also provide the necessary validation for the models utilizing only computational data. We also note that there are caveats in the experimental data, which naturally translate into limitations on the predictive accuracy of ML models. In most cases, only the major forms of magnetic ordering — i.e., paramagnetism (PM), FM, or AFM (denoted as classes in classification models) — are reported, while information about the specific types of AFM or PM, or the orientation of magnetic moments with respect to crystallographic axes is not given. The flowchart shown in Fig. 4.7, outlines the main stages involved in the development of the ML models utilized in this study. These stages include compilation and curation of appropriate datasets, performing data analysis with standard data mining tools, construction of the models and their following
internal and external validation.

In order to predict the magnetic moment sizes, we have developed our ML-based framework focusing on *regression* type algorithms. Five different algorithms [240], including LR, LASSO, KRR RFR and SVMR were chosen systematically to evaluate the performance of the models. The hyperparameters utilized to build these models are given in the Appendix (Chapter 6). We have also utilized random forest classification (RFC) algorithm to predict the probabilities of each compound exhibiting PM, FM or AFM ordering. We note that the choice of these algorithms [241] is driven by transferability, simplicity of implementation, as well as compliance with the best practices in this area of research.

The main findings of our study are summarized as follows: (a) ML frameworks employing regression and classification algorithms can be utilized to predict magnetic moment size and ordering in uranium-based binary compounds. (b) Various descriptor sets, depending on the availability of data sources (both computational and experimental), can be designed to construct families of meaningful ML models. (c) While ML models can identify the most important descriptors pertinent to the endpoints, complementary data analysis can establish quantitative structure-property relationships (e.g., cut-off values of lattice parameters for a compound to exhibit PM, FM or AFM ordering) present in such materials. (d) It is often possible to identify physical phenomena underpinning the results obtained from ML models (e.g., exchange interactions that lead to particular types of magnetic behavior).
Figure 4.7: Flowchart outlining the main development stages involved in construction and validation of ML models for predicting cation magnetic moment size and magnetic ordering in actinide-based binary systems. Primary stages are shown as grey rectangles, while any necessary secondary stages are represented by white rectangles. Some of the diagram elements introduced here are analyzed in further detail in Fig. 4.9 and accompanying text.
Figure 4.8: Structural models of 12 uranium-based binary compounds that were assumed in DFT computations for creating Dataset I. The AFM I configurations are shown using a vector representation, where directions of magnetic moments on U cations are alternating between consecutive uranium layers and their orientation is along the crystal c axis.

4.2.2 Methods

4.2.2.1 Datasets

First-principles calculations of average spin and orbital moments were performed using the PAW method implemented in the VASP [84, 85]. GGA was adopted to represent the exchange and correlation interactions, with the GGA+\(U_{\text{eff}}\) [242] approach utilized to capture the strongly correlated nature of the 5\(f\) electrons. All computations were carried out with a 500 eV plane-wave cutoff energy using tetrahedron method with Blochl corrections with appropriate MP [74] \(k\)-point meshes, which produced well converged
For the construction of Dataset I, which is built only on the data extracted from the DFT simulations, the following twelve uranium-based binary compounds were utilized: UO$_2$, U$_3$O$_8$, UO$_3$, UN, UC, UP, UP$_2$, U$_3$P$_4$, UAs, UBi$_2$, USB$_2$ and UCl$_3$. The magnetic structures of these compounds are well documented in the literature [243–246], which is the primary motivation behind choosing them to build Dataset I. Geometrical structures with AFM-I configuration for all of these compounds are shown in Fig. 4.8. For each compound, initial lattice parameters (listed in Table 4.5) and ionic positions were obtained from ICSD [247], after which eight individual variants were created by varying the Hubbard parameter $U_{\text{eff}}$ between 0 and 6 eV in 2 eV increments in the presence or absence of spin-orbit coupling. $U_{\text{eff}}$ values from the same range have been used previously in a number of DFT-based investigations of actinide compounds. Electronic and magnetic properties for each of the eight variants were evaluated both for the ICSD provided structural parameters and after optimization, which included relaxing the unit-cell shape and volume to stresses below 0.1 kbar and all the ionic positions until the associated Hellman-Feynman forces were below $10^{-3}$ eV/Å. Utilization of data generated using structural parameters obtained from both ICSD and DFT computations allowed us to incorporate information of varying fidelity [248] levels into our ML models, which is in turn helpful for consistent evaluation and improvement of their predictive accuracy. We note that structural configurations obtained from DFT computations belong to the same space groups as those reported in ICSD. Although it is possible to create multiple
other structures by directly varying lattice parameters and internal positions of atoms, the descriptor space of Dataset I will not be significantly altered. Therefore, inclusion of such additional entries into Dataset I will not help improve predictive capabilities of the resulting ML models.

For all computations involved in generating Dataset I, we considered a specific type of AFM ordering, namely AFM I, where the directions of magnetic moments on U cations are alternating between consecutive uranium layers and their orientation is along the crystal $c$ axis. We restricted ourselves to studying just one magnetic configuration type since our primary interest is in estimating moment sizes only — and these do not change significantly if other configurations (e.g., out-of-plane spin orientation) are selected instead.

The chosen initial AFM configurations for all 12 compounds are in accordance with those reported in the literature [243–246]. For $U_3P_4$, a FM configuration has lower energy, as compared to that in AFM I. This is confirmed by DFT+$U_{\text{eff}}$ (eV) computations showing the energy difference of $\sim 3.26$ eV between these two magnetic configurations. The energy trade-offs between choosing in-plane vs. out-of-plane magnetic orientations for all 12 compounds are listed in Table 4.6.

We report nominal differences ($< 0.3 \, \mu_B$) between these two AFM configurations as compared to average spin (1.64 $\mu_B$) and orbital (2.82 $\mu_B$) moment sizes for all of the chosen compounds. Overall, Dataset I was built solely using DFT-simulations and comprises 16 variants for every compound, for a total of 192 entries.
Table 4.5: Structural parameters and space groups of 12 compounds from Dataset I. All of these compounds have the same AFM I (self-consistently converged) configuration.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Lattice Parameters (Å)</th>
<th>Space Group</th>
<th>Hubb $U_{\text{eff}}$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UO$_2$</td>
<td>$a = b = c = 5.4682$</td>
<td>225, $Fm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>U$_3$O$_8$</td>
<td>$a = 6.7039$, $b = 11.9499$, $c = 4.1420$</td>
<td>21, $C222$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UO$_3$</td>
<td>$a = 8.3299$, $b = 4.1649$, $c = 4.1649$</td>
<td>221, $Pm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UN</td>
<td>$a = b = c = 4.4889$</td>
<td>225, $Fm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UC</td>
<td>$a = b = c = 4.9597$</td>
<td>225, $Fm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UP</td>
<td>$a = b = c = 5.5869$</td>
<td>225, $Fm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UAs</td>
<td>$a = b = c = 5.7767$</td>
<td>225, $Fm\bar{3}m$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UP$_2$</td>
<td>$a = b = 3.8099$, $c = 7.7639$</td>
<td>129, $P4/nmm$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UBi$_2$</td>
<td>$a = b = 4.4450$, $c = 8.9079$</td>
<td>129, $P4/nmm$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>USb$_2$</td>
<td>$a = b = 4.2719$, $c = 8.7410$</td>
<td>129, $P4/nmm$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>U$_3$P$_4$</td>
<td>$a = b = c = 8.2119$</td>
<td>220, $I\bar{4}3d$</td>
<td>0, 2, 4, 6</td>
</tr>
<tr>
<td>UCl$_3$</td>
<td>$a = b = 7.4429$, $c = 4.3210$</td>
<td>176, $P6_3/m$</td>
<td>0, 2, 4, 6</td>
</tr>
</tbody>
</table>

Dataset II was constructed by curating the results of 737 experimental reports of standard quality on uranium-based binary compounds, as found in the ICSD [247]. Only structures stable at low temperature were considered, while data on any metastable high-temperature configurations was discarded. After the removal of duplicate entries, 223 data points including information on magnetic properties (cation moment size and ordering) remained in the dataset. Dataset II contains the following unique compounds: UAl$_2$, UAl$_3$, UCo$_2$, UCo$_3$, UCo$_5$, UFe$_6$, UGa$_2$, UGa$_3$, UIn$_3$, UIr, UNi$_2$, UNi$_5$, UNi$_6$, UOs$_2$, UPd$_3$, UPt$_5$, UPt, URh$_3$, USi$_3$, USn$_3$, UTl$_3$, U$_2$N$_3$, U$_2$S$_3$, U$_2$Se$_3$, U$_2$Te$_3$, U$_2$Te$_5$, U$_2$Ti, U$_3$As$_4$, U$_3$Bi$_4$, U$_3$P$_4$, U$_3$Pt$_4$, U$_3$S$_5$, U$_3$Sb$_4$, U$_3$Se$_4$, U$_3$Se$_5$, U$_3$Si, U$_3$Si$_2$, U$_3$Si$_5$, U$_3$Te$_5$, U$_5$Ge$_3$, U$_5$Ge$_4$. 
Table 4.6: Differences in energy and moment sizes between in-plane and out-of-plane moment orientations for all 12 compounds present in Dataset I. Spin-orbit coupling and $U_{\text{eff}} eV$ were used in these computations.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$\Delta E$ (eV)</th>
<th>$\Delta \mu_{\text{spin}} (\mu_B)$</th>
<th>$\Delta \mu_{\text{orbit}} (\mu_B)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UO$_2$</td>
<td>0.032</td>
<td>0.001</td>
<td>-0.009</td>
</tr>
<tr>
<td>U$_3$O$_8$</td>
<td>0.044</td>
<td>0.071</td>
<td>-0.011</td>
</tr>
<tr>
<td>UO$_3$</td>
<td>0.096</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UN</td>
<td>0.007</td>
<td>0.15</td>
<td>-0.037</td>
</tr>
<tr>
<td>UC</td>
<td>0.069</td>
<td>0.054</td>
<td>-0.074</td>
</tr>
<tr>
<td>UP</td>
<td>-0.003</td>
<td>-0.090</td>
<td>-0.068</td>
</tr>
<tr>
<td>UAs</td>
<td>-0.002</td>
<td>-0.006</td>
<td>-0.02</td>
</tr>
<tr>
<td>UP$_2$</td>
<td>-0.206</td>
<td>0.008</td>
<td>-0.036</td>
</tr>
<tr>
<td>UBi$_2$</td>
<td>0.032</td>
<td>0.029</td>
<td>0.023</td>
</tr>
<tr>
<td>USb$_2$</td>
<td>0.024</td>
<td>0.028</td>
<td>0.038</td>
</tr>
<tr>
<td>U$_3$P$_4$</td>
<td>0.067</td>
<td>-0.011</td>
<td>0.087</td>
</tr>
<tr>
<td>UCl$_3$</td>
<td>0.279</td>
<td>-0.009</td>
<td>-0.011</td>
</tr>
</tbody>
</table>

UAs, UC, UCl$_3$, UC$_5$, UD$_3$, UGa$_2$, UGe$_2$, UH$_3$, UI$_3$, UIr, UN, UP, UP$_2$, US, US$_2$, US$_3$, USb$_2$, USE, USE$_2$, USE$_3$, USi$_2$, USn$_3$, UTe, UTe$_2$, UTe$_3$. These compounds were also categorized into three classes according to the nature of the reported magnetic ordering: (1) PM, i.e., compounds with local magnetic moment but no long range order present, (2) FM, i.e., compounds with magnetic spins aligned in the same directions and (3) AFM A-type or C-type.

The sizes of both datasets before and after curation, as well as the filtering criteria for Dataset II are shown in the top part of the ML model development flowchart in Fig. 4.9.
Figure 4.9: Flowchart showing the stages of development of ML models to predict both moment size and ordering.
For each dataset, 20 entries were kept aside for internal validations and the rest (172 for Dataset I and 203 for Dataset II) were used for training (90%) and testing (10%) of ML models.

### 4.2.2.2 Descriptors

For Dataset I, the following eight primary descriptors were considered: lattice parameters \( \text{magnetic unit cell parameters} / \sqrt{\text{number of actinide elements}} \) \( alatt, blatt \) and \( clatt \) (Å), atomic volume \( \text{magnetic unit cell lattice parameters} / \text{number of actinide elements} \) \( \text{volume (Å}^3) \), Hubbard parameter \( U_{\text{eff}} \) (eV), spin-orbit coupling strength SOC (eV), cation 5f-subshell occupation number (number of total valence electrons of actinide element – valence of anion) \( N_{\text{occ}}(5f) \) and system Fermi energy level \( E_{\text{F}} \) (eV). For each primary descriptor \( x \), additional compound descriptors were generated using 10 prototypical functions, namely, \( x^2, x^3, \exp(x), \sin(x), \cos(x), \tan(x), \sinh(x), \cosh(x), \tanh(x), \) and \( \ln(x) \), to allow for possible non-linearities in the connections between the descriptor and endpoint properties.

The descriptor space for Dataset II contains all structural parameters (as defined above), number of formula units and \( N_{\text{occ}}(5f) \) — all extracted from the respective experimental reports. Furthermore, for every entry in Dataset II, a matrix representation called Orbital field matrix (OFM) [249, 250], as implemented in a Python library [251], was computed using distances between coordinating atoms, valence shells and Voronoi polyhedra weights, which provided information on the chemical environment of each atom in the unit cell.
4. Traditional ML-algorithms based learning

The OFM elements are defined as following [249, 250]:

\[ X_{ij}^p = \sum_{k=1}^{n_p} o_i^p o_j^k \frac{\theta_k^p}{\theta_{\text{max}}^p} \zeta(r_{pk}), \quad (4.1) \]

where \( i, j \in D = (s^1, s^2, \ldots, f^{14}) \) for central \( i \) and coordinating \( j \) orbitals, respectively; \( o_i^p \) and \( o_j^k \) are elements of one-hot vectors \((i, j)\) of the electronic configuration \( p \) and neighboring atom indexed by \( k \). The weight of the atom \( k \) in the coordination of the central atom at site \( p \) is given by \( \frac{\theta_k^p}{\theta_{\text{max}}^p} \), where \( \theta_k^p \) is the solid angle determined from the respective Voronoi polyhedron. The number of nearest-neighbor atoms surrounding atom site \( p \) is \( n_p \). The size and distance-dependent weight function is also included as \( \zeta(r_{pk}) \).

4.2.2.3 Data Analytics

Prior to the development of ML models, both datasets were analyzed using standard data analytics techniques to determine the presence of any relationships between the structural descriptors and the targeted magnetic properties. Any inferences drawn from this analysis may provide useful insight for choosing appropriate descriptors space and algorithms for the ML model development. The primary descriptor sets for Dataset I were subjected to a Pearson correlation filter to remove features that exhibit a high correlation with the other descriptors within the same set, as shown in Fig. 4.10. The same approach was also applied to the combined set of primary and compound descriptors. Additionally, for Dataset I we employed a conditional inference procedure with Bonferroni-corrected significance (p-value < 0.05) value, used as the splitting criteria (stopping rules)
Figure 4.10: Correlation matrices representing Pearson correlation coefficients for primary descriptors applicable to Dataset I for (a) spin moment size and (b) orbital moment size endpoints. The primary descriptor space for Dataset I consists of eight features: lattice parameters $a_{\text{lat}}$, $b_{\text{lat}}$ and $c_{\text{lat}}$, and volume, Hubbard parameter $U_{\text{eff}}$, spin-orbit coupling strength (SOC), cation 5$f$-subshell occupation number $N_{\text{occ}}(5f)$ and system Fermi energy level $E_{\text{F}}$. (We note that the $E_{\text{F}}$ descriptor was not used in the ML model construction, as described later in Section 4.2.3). Learning curves for two ML models predicting (c) spin moment size and (d) orbital moment size constructed using the RFR algorithm. $E_{\text{Train}}$ and $E_{\text{Test}}$ refer to average training and test set root mean square errors. The average % mean absolute test errors in predicting spin (c) and orbital (d) moment sizes are 14%, 17%.
4. Traditional ML-algorithms based learning

Figure 4.11: Dataset II was divided into two sets: below median value (grey), above median value (orange). The difference between these two sets for all experimental descriptors is represented by the blue bar. The considered features were (a-c) lattice parameters (Å) in all directions, (d) cell volume (Å³), (e) formula units (FU). The grouping was done based on the respective medians and type of magnetic ordering reported. The median values for these features are (a) 5.588 Å, (b) 5.5 Å, (c) 5.62 Å, (d) 191.85 Å³ and (e) 4. The ML-based prediction with 76% average accuracy on the test set for Dataset II is given by the (f) confusion matrix. Here, the highlighted boxes denote the number of compounds for which the magnetic ordering class was predicted correctly.

for each node while constructing trees as implemented in R version 3.4.2 via CTree algorithm [91]. The splitting process was continued recursively throughout the whole Dataset I.

We utilized all 223 entries in Dataset II to perform median analysis. The dataset was divided into two subsets (below and above the median value) based on the median value
for each descriptor. For every descriptor, magnetic ordering class was assigned to the respective entries belonging to the two resultant subsets. Bar charts shown in Fig. 4.11(a-e) report the differences between the mean values of the descriptors. Larger differences observed for any particular descriptor are likely to produce an increased variance in the endpoint when that descriptor is used for construction of ML models.

4.2.2.4 Algorithms

Five regression algorithms, including LR, LASSO, KRR, RFR, SVMR as implemented in Scikit-learn Python [93] version 3.5.2 were employed to construct ML models from Dataset I. Regression-based algorithms were chosen over classification-based ones, as the endpoint of interest (magnetic moment size) was computed numerically for each entry in respective datasets. The RFC classification algorithm was also utilized to construct ML models for predicting the magnetic ordering endpoint.

4.2.2.5 ML model development and validation

One of the main concerns of conventional ML model development is selection of the dataset size. Here, for both datasets, the number of entries was restricted to less than 230 entries, which may potentially limit the accuracy of the resulting models. The standard deviations in the predicted values for the test sets may be as large as 5-10%, making any comparisons originating from a single set of predictions misleading, where two such cases are compared). In order to avoid any statistical bias, we built learning curves, such as ones
shown in Fig. 4.10, by varying the training set size in order to estimate the performance of all the developed models. For each point on the learning curve, the average RMSE is calculated using 1000 randomly generated (by sampling with replacement) training and test set evaluations. The average RMSE for a training set size \( N \) is denoted \( E_{\text{Train}}(N) \), whereas for the corresponding test set it is denoted \( E_{\text{Test}}(N) \), however, in this case the size of the set is now the total number of points minus \( N \). The test set RMSE provides the expected error for a given model, while the difference between \( E_{\text{Test}}(N) \) and \( E_{\text{Train}}(N) \) is an estimation of how much variance or overfitting the model contains. We have used confusion (error) matrices [93] to evaluate the performance of the RF classification models. For the three classes, such as AFM, FM, and PM, these matrices show the predicted class of each compound together with its true class. Diagonal elements of a confusion matrix represent the number of compounds for which the predicted classes match the true classes.

Both models predicting the moment size and ordering were tested using the internal validation sets, i.e., 20 entries that were kept aside on the model development stage [Fig. 4.9]. Moreover, we also applied these models to three external sets of actinide-based binary and ternary compounds, for which information on either moment size or ordering is not widely available in literature. Validations on datasets composed of materials that are dissimilar to those used in model development provided more stringent testing of robustness and transferability of our ML models.

We note that the SOC descriptor (representing the strength of such coupling determined by the DFT computations for Dataset I) was unavailable for the compounds present
in the validation set. As an alternative way to include this interaction in our models we considered SOC as being present (1) or absent (0) instead of including its strength value. The $U_{\text{eff}}$ could also be varied manually in the descriptor space depending on its estimated value (eV) to capture the strong correlation effect among the $f$-orbital electrons. In addition, valence electron numbers for other atoms and OFM, as applicable to one of the external validation sets consisting of uranium-based ternary compounds, were also included in the feature space for testing models built on Datasets I and II.

### 4.2.3 Results and Discussion

Here, we discuss the results obtained from the preliminary data analysis and ML models applied to internal and external validation sets of Datasets I and II. The discussion proceeds in a sequential order, with each primary step highlighted throughout the section.

From the results of the correlation analysis presented in Fig. 4.10(a) for the set of primary features, it is evident that volume, $N_{\text{occ}}(5f)$ and $En_F$ descriptors have large Pearson coefficients with respect to the spin moment size endpoint, suggesting that these features should be included in the descriptor space for building ML models to predict spin moment size. However, since the $N_{\text{occ}}(5f)$ and $En_F$ descriptors are highly correlated to each other, we kept only the former as a primary descriptor. We note that the $En_F$ descriptor obtained from the DFT computations is relative and can be placed anywhere in the energy
band gap between the occupied and unoccupied states. Therefore, the $N_{\text{occ}}(5f)$ feature is more appropriate as a descriptor containing information related to the electron density of states, which is apparently important for predicting magnetic moment sizes. The SOC descriptor has the highest correlation coefficient with respect to the orbital moment size endpoint in the matrix, as shown in Fig. 4.10(b). This is expected for the primary feature that accounts for strong SOC originating from $5f$ shell electrons. This analysis, extended to space of primary and secondary features, allowed us to shortlist 61 features by excluding highly-correlated ones (correlation factor > 0.85). Next, the conditional inference procedure applied to Dataset I shows that $a_{\text{lat}}, \text{volume}, U_{\text{eff}}$ and SOC are the top features capable of grouping the data well for spin and orbital moment size as endpoints, respectively.

**Median analysis:** as explained in Section 4.2.2.3, we have used a different approach to analyze Dataset II by determining medians and differences in average values of the features. Based on the median values of the five features, Dataset II was divided into subgroups, which was followed by assignment of the corresponding magnetic ordering and calculation of the difference (variance) in average feature values for the subgroups, as shown in Fig. 4.11(a-e). The median values for each of these experimental descriptors ($a_{\text{lat}}$, $c_{\text{lat}}$, $a_{\text{lat}}$, number of formula units) are 5.588 Å, 5.5 Å, 5.62 Å, 191.85 Å³ and 4 respectively. There is a total of 112 entries in the subgroup where $a_{\text{lat}} \geq$ median value of $a_{\text{lat}}$, out of which 24 are AFM, 44 are FM and 44 are PM. For the similar subgroups formed by other features, such as $b_{\text{lat}}, c_{\text{lat}}, \text{volume}, \text{number of formula units}$, the number
of entries are: (26 - AFM, 50 - FM, 36 - PM), (30 - AFM, 46 - FM, 36 - PM), (24 - AFM, 48 - FM, 40 - PM) and (42 - AFM, 70 - FM, 51 - PM), respectively.

Both conditional inference tree and median-variance methods suggest that the decision tree type of algorithm may have better performance compared to other regression algorithms, if used to build ML models for predicting moment size. Overall, these investigations performed using standard data analytics techniques are useful for cultivating some advance knowledge about the available data and identification of features important to predicting endpoints, as well as any inconsistencies present in the datasets.

The results produced by the ML models for predicting the moment size are presented as learning curves in Fig. 4.10(c,d). For both cases, the RFR algorithm outperforms the other algorithms significantly, as shown in both Fig. 4.10(c,d) and Table 4.7. The average spin moment size of compounds used for training these models is 1.64 \( \mu_B \). For the orbital moment size, the average is 2.82 \( \mu_B \), counting only the training data points computed by including the SOC. The total moment size can be obtained using the vector sum of both spin and orbital moments, pointed in the opposite direction to each other due to Hund’s rule for \( f \)-electron shell that is less than half filled. The average \( E_{\text{Test}} \) values in predicting spin and orbital moment size are 0.17 \( \mu_B \) and 0.19 \( \mu_B \), respectively, as mentioned in Table 4.7. In both cases, the RFR algorithm overfits the data by approximately 4%. On the other hand, the LASSO, KRR and SVMR algorithms display minimal overfitting but plateau at higher \( E_{\text{Test}}(N) \), as shown by comparative learning curves. Furthermore, all of the ML models are performance-optimized based on bias-variance trade-offs, i.e., these
Table 4.7: Test set error $E_{\text{Test}}$ and difference (overfitting) between the training $E_{\text{Train}}$ and test set errors for the training set size $N = 50$ and the largest training set size $N_{\text{max}}$. The predicted endpoints are magnitudes of spin and orbital magnetic moments in $\mu_B$.

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Algorithm</th>
<th>$E_{\text{Test}}$ (50)</th>
<th>$E_{\text{Test}} - E_{\text{Train}}(N_{\text{max}})$</th>
<th>$E_{\text{Test}}(50)$</th>
<th>$E_{\text{Train}}(N_{\text{max}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spin moment</td>
<td>LR</td>
<td>0.33</td>
<td>0.17</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.30</td>
<td>0.08</td>
<td>0.30</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>KRR</td>
<td>0.30</td>
<td>0.28</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.84</td>
<td>0.04</td>
<td>0.84</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>RFR</td>
<td>0.32</td>
<td>0.13</td>
<td>0.17</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orbital moment</th>
<th>Algorithm</th>
<th>$E_{\text{Test}}$ (50)</th>
<th>$E_{\text{Test}} - E_{\text{Train}}(N_{\text{max}})$</th>
<th>$E_{\text{Test}}(50)$</th>
<th>$E_{\text{Train}}(N_{\text{max}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
<td>1.46</td>
<td>1.66</td>
<td>1.46</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.90</td>
<td>0.14</td>
<td>0.80</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>KRR</td>
<td>0.96</td>
<td>0.30</td>
<td>0.96</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>1.07</td>
<td>0.40</td>
<td>1.01</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>RFR</td>
<td>0.41</td>
<td>0.22</td>
<td>0.19</td>
<td>0.03</td>
</tr>
</tbody>
</table>
models always yield the optimized moment size, when used for any other material including those not present in the training set.

We have used confusion (error) matrices [93] to evaluate the performance of the RFC classification models. For the three classes, such as AFM, FM and PM, these matrices show the predicted class of each compound along with its true class. Diagonal elements of a confusion matrix represent the number of compounds for which the true and predicted classes match. The average accuracy in predicting magnetic ordering for the test set of Dataset II is 76%, as displayed in Fig. 4.11(f). The OFM representation plays a key role in significantly improving the model performance (accuracy improved by ∼15%) in predicting magnetic ordering by including information on the valence shells and local coordination environment of the system. For both the moment size and ordering endpoints, the most important features (structural parameters and \( f \)-subshell occupation numbers) identified by the ML models based on the RFR and RFC algorithms are the same as those found earlier by the data analytics techniques.

We have also compared the average nearest neighbor distances (\( d_{U-U} \)) for every entry in Dataset II to the Hill limit [252] that provides restrictions, under which magnetic ordering occurs in actinide systems as shown in Fig. 4.12. For \( d_{U-U} < 3.4 \, \text{Å} \), intermetallic uranium compounds are PM at low temperatures. Due to large overlap of \( 5f \) orbitals, compounds with \( d_{U-U} > 3.6 \, \text{Å} \) generally exhibit long-range FM or AFM ordering, localized at uranium ions. On the other hand, for smaller \( d_{U-U} \), itinerant \( f \) electron behavior is more dominant leading to temperature-independent PM ordering. Our results are in
reasonable agreement with the above mentioned Hill limit for uranium ions to exhibit specific type of ordering. This observation is important to establish a physical significance of lattice parameters, which are identified as features critical for predicting both the moment size and ordering. The average lattice parameters obtained from Dataset I are comparable within 15% to entries present in Dataset II that reportedly exhibit AFM ordering at low temperatures. This, along with results from median-variance analysis performed on Dataset II, also provides quantitative measure for structural parameters to observe a specific type of magnetic ordering in high likelihood. For example, a uranium-based binary compound with $a_{\text{latt}} \geq 5.58\text{Å}$, is more likely to exhibit AFM ordering at low temperatures.

Three of the models were tested first on the internal validation sets kept aside within Datasets I and II. We employed an approach comparable to the learning curves by re-
Figure 4.13: ML model predictions of (a) spin and (b) orbital moment size for the internal validation set from Dataset I. A comparison between experimentally reported and predicted total moment size is presented in panel (c) for compounds in the validation set for which the moment sizes are reported in the literature. The average RMSEs for the prediction of total moment size for these compounds are 0.32 $\mu_B$ and 0.35 $\mu_B$ for $U_{\text{eff}} = 4$ and 6 eV, respectively. Magnetic ordering of compounds present in the internal validation set was predicted utilizing a classification model constructed on Dataset II. The confusion matrix shown in panel (d) aggregates the true and predicted ordering classes for each entry with 70.2% average accuracy.
porting the average moment size and ordering for each entry, as obtained by averaging over 1000 ML model predictions. Fig. 4.13(a,b) shows predictions made on the internal validation set acquired using Dataset I. The average RMSEs for the spin and orbital moment size predictions on the internal validation sets (for $U_{\text{eff}} = 4$ eV) are 0.20 $\mu_B$ and 0.25 $\mu_B$, respectively. For compounds, such as UN, USB$_2$, UO$_2$, UBi$_2$ and UP, total moment sizes are available in the literature, and can be used to compare the ML model predictions for $U_{\text{eff}} = 4$ and 6 eV, as shown in Fig. 4.13(c). The average RMSEs for the prediction of the total moment size for these compounds are 0.32 $\mu_B$ and 0.35 $\mu_B$ for $U_{\text{eff}} = 4$ and 6 eV, respectively. This analysis also highlights the dependence of moment size on $U_{\text{eff}}$ that is captured by the ML models. The average prediction accuracy obtained using RFC for the internal validation set of Dataset II is 70.2%. The numbers of entries belonging to the PM, FM and AFM classes, for which the predictions have matched with the observed orderings are 9, 2 and 2, respectively, as shown in Fig. 4.13(d).

Finally, to assess the performance and transferability of these models to other actinide systems, we have compiled three external validation sets, first two containing uranium-based and neptunium-based binary and third one containing uranium-based ternary compounds [253]. The results of these predictions using the same averaging technique as applied before to the internal validation sets are shown in Fig. 4.14. We note that although predictions of the moment size cannot be verified for these compounds due to the scarcity of experimental information, predictions of magnetic ordering that classify them as AFM, FM or PM are comparable with those reported in the literature [253] without accounting
for the exact spin textures.

The *external* Set I of binary compounds includes 34 different uranium-based crystals: $U_2C_3$, $U_3As_4$, $U_3Bi_4$, $U_3Sb_4$, $UAl_2$, $UAi_3$, $UAi_2$, $UB_2$, $UB_4$, $UBi$, $UCo_2$, $UFe_2$, $UGa_3$, $UGe_2$, $UGe_3$, $UIn_3$, $UIr_2$, $UIr_3$, $UIr$, $UMn_2$, $UNi_2$, $UPb_3$, $UPd_3$, $UPt_2$, $UPt_5$, $URh_3$, $US$, $USb$, $USe$, $USi_3$, $USn_3$, $UTE$ and $UTl_3$. These are not present in either of the Datasets (I & II) used in model development reported above. The top 5 most common structure types in this list belong to cubic crystal family with space group numbers 221, 220, 225, 227 and 191. There are 8 compounds that exhibit AFM, 10 with FM and the rest have PM ordering at low temperatures. For ordering, the average (based on the ML model built on Dataset II) accuracy is 70.1% as predicted for all 34 compounds, shown in Fig. 4.14(a,b). While the confusion matrix provides the number of entries with respective true and predicted classes, the bar-chart reports the probability of each compound belonging to either PM, FM or AFM, summing up to 1 or 100%.

*External* Set II has a total of 43 entries of 35 unique neptunium-based compounds: $NpAl_2$, $NpAl_3$, $NpAl_4$, $NpGa_2$, $NpGa_3$, $NpIn_3$, $NpIr_2$, $NpIr_3$, $NpNi_2$, $NpNi_3$, $NpO_5$, $NpSe_5$, $NpS_5$, $NpS_5$, $NpS_5$, $NpAs$, $NpAs_2$, $NpB_2$, $NpC$, $NpCo$, $NpFe_2$, $NpGe_3$, $NpIn_3$, $NpMn_2$, $NpN$, $NpN_2$, $NpNi_2$, $NpO_2$, $NpOs_2$, $NpP$, $NpPd_3$, $NpS$, $NpSb$, $NpSb_2$, $NpSi_2$, $NpSi_3$ and $NpSn_3$. All 43 entries are considered for predictions of magnetic ordering using the ML model built on Dataset II. As shown in Fig. 4.14(c), there are total of 18 AFM, 23 FM and 2 PM compounds present in this set. Out of all AFM compounds, 14 are predicted with the highest probability belonging to the true class, whereas for FM and PM, there are total
Figure 4.14: ML model predictions of magnetic ordering for three external validation sets including uranium-based binary (row 1), neptunium-based binary (row 2) and uranium-based ternary compounds (row 3) represented by confusion matrices (a,c,e) and charts (b,d,f) showing probabilities for each compound to belong to a specific ordering class. The average prediction accuracies of these models are 70.1%, 67.5% and 68.9%, respectively.
of 16 compounds with correct classification of ordering prediction with average accuracy of 67.5%. The most common structure type among these compounds is rocksalt cubic, followed by Laves phase cubic and Auricupride. These three structure types are also common among compounds in Datasets I and II. Hence, this suggests that the developed ML models may be capable of predicting the endpoints with similar accuracy as reported earlier in the section.

External Set III containing ternary compounds has 136 different entries, out of which 57 exhibit AFM ordering, 39 show FM ordering and the rest are paramagnetic at low temperatures. The list of compounds present in external set III is the following: U₂Co₂In, U₂Co₂Sn, U₂Co₃Si₅, U₂Fe₂Sn, U₂Ir₂Sn, U₂Mo₂Ge₄, U₂Mo₄Si₄, U₂Nb₃Ge₄, U₂Ni₂In, U₂Ni₂Sn, U₂Pd₂In, U₂Pd₂Sn₂, U₂Pt₂In, U₂Pt₂Sn, U₂PtC₂, U₂Rh₂Sn, U₂RhIn₈, U₂Ru₂Sn, U₂Ta₃Ge₄, U₂Wi₃Si₄, U₃Al₂Si₃, U₃Au₃Sn₄, U₃Co₃Sb₄, U₃Cu₃Sb₄, U₃Cu₅Sn₄, U₃Cu₄Ge₄, U₃Ir₃Sb₄, U₃Ni₃Sb₄, U₃Ni₃Sn₄, U₃Ni₄Si₄, U₃Pd₃Sb₄, U₃Pt₃Sb₄, U₃Pt₃Sn₄, U₃Rh₃Sb₄, U₃Rh₄Sn₁₃, U₃Ru₄Al₁₂, U₄Os₇Ge₆, U₄Re₇Si₆, U₄Ru₇Ge₆, U₄Tc₇Ge₆, U₄Tc₇Si₆, UAsSe, UAsTe, UAu₂Al, UAu₂In, UAu₂Sn, UAu₂Al, UAu₃Ga, UAu₃Ge, UAu₃Si, UAu₂Sn, UCo₂Ge₂, UCo₂P₂, UCoAs₂, UCoGa₅, UCoGa, UCoGe, UCoP₂, UCoSi, UCoSn, UCoSn, UCr₂Si₂, UCrC₂, UCu₂As₂, UCu₂Ge₂, UCu₂P₂, UCu₂Si₂, UCu₂Sn, UCu₂Bi₂, UCuGa, UCuGe, UCuP₂, UCuSb₂, UCuSi, UCuSn, UFe₂Ge₂, UFe₂Si₂, UFe₂Al, UFe₃As₂, UFeGa₅, UFeGa, UFeGe, UFeSi, UIr₂Ge₂, UIr₂Si₂, UIrAl, UIrGe, UIrSi₃, UIrSi, UIrSn, UMn₂Ge₂, UMn₂Si₂, UNi₁₆As₂, UNi₂Al₃, UNi₂Ga, UNi₂Ge₂, UNi₂Si₂, UNi₂Sn, UNi₃Al, UNi₃Ga, UNi₃Ga, UNi₅Ga, UNi₅Ge, UNi₅Sb₂, UNi₅Si, UNi₅Sn, UOsGa₅, UPd₂Al₃, UPd₂Ga, UPd₂Si₂, UPd₂Sn, UPdGa₅, UPdGe, UPdIn,
UPdSb, UPdSi, UPdSn, UPt$_2$Si$_2$, UPt$_4$Au, UPtAl, UPtGa$_5$, UPtGe, UPtIn, UPtSi, UPtSn, URh$_2$Ge$_2$, URhAl, URhGa$_5$, URhGe, URhIn$_5$, URhSi, URhSn, URu$_2$P$_{1.894}$, URu$_2$Si$_2$, URu$_2$Si$_2$, URu$_4$B$_4$, URuAl, URuGa$_5$, URuSb, URuSn.

These compounds commonly belong to families of orthorhombic, tetragonal, hexagonal and cubic crystal systems, with the five most common space groups being 62, 139, 123, 127 and 189. For the determination of ordering, the average (based on the ML model built on Dataset II) accuracy is 68.9% as predicted for all 136 compounds, shown in Fig. 4.14(e,f).

Overall, the ML-based models built on Datasets I and II are capable of delivering reasonable predictions of moment size and ordering for actinide-based binary and ternary compounds.

We note that a regression-based ML framework (such as models based on the RFR algorithm) is equally suited to predict magnetic ordering as the classification-based one discussed above. From Fig. 4.11(a-e), it is evident that the experimental descriptors are capable of dividing Dataset II according to each class of ordering and moment size. The difference in the number of compounds exhibiting a specific type of ordering is varied according to median of the descriptor values, meaning larger difference for a particular descriptor should lead to a greater variance when used in decision-tree type algorithms (such as RFR). This was one of the primary reasons for evaluating such framework for ordering predictions. Instead of partitioning the compounds into three classes and predicting those, Dataset II can utilize labels 1, 2 and 3 to represent PM, FM and AFM or-
4. Traditional ML-algorithms based learning

dering. In that case, a score corresponding to each entry can be predicted with a certain RMSE. Later, a clustering algorithm can be employed to partition these compounds into PM, FM or AFM classes. The average RMSE for predicting the ordering with this scheme is 0.12, which is comparable to the accuracy achieved using the RFC algorithm. The OFM representation plays a key role in significantly improving the model performance (RMSE reduced by \(\sim 30\%\)) in predicting magnetic ordering, in the same fashion as already discussed above. Our regression algorithm based ML framework can predict numerical scores for partitioning compounds into AFM, FM and PM ordering classes with an average RMSE of 0.15 for structures in the internal validation set of Dataset II. The average RMSEs are 0.23 (based on the ML model built on Dataset II), 0.14 and 0.24, respectively, as predicted for all compounds listed in external validation sets I, II and III.

These prediction accuracies are comparable with those obtained using the classification approach. Moreover, since classification algorithms treat separate classes as discrete rather than continuous entities, they cannot provide further information on possible phase transitions occurring at low temperatures. Instead, regression algorithm based predictions can deliver non-integer scores (e.g., an ordering of 1.5), suggesting that the compound of interest may be near the boundary between the PM and FM phases. UTe\(_2\) compound, present in the internal validation set of Dataset II, may serve as an example. The classification model assigns it to the PM class. However, the regression algorithm gives it an ordering score of 1.46, which falls between PM and FM, indicating of the existence of FM fluctuations at low temperatures. Recently, this compound has been indeed proposed
to be at the verge of a FM phase and have a conventional spin-triplet superconducting pairing [254, 255]. In addition, as inferred from the classification model, a representative heavy-fermion magnet UCu$_5$ with a Neél temperature of $\sim 12$K [256] belongs to the AFM class, while the regression approach gives it a score of 2.63, which may be indicative of possible phase transitions between multiple ordering phases. Given that the temperature dependence of magnetic properties in heavy fermion compounds is a known fact, additional physical significance can be attributed to the fractional scores obtained from such ML models.

### 4.2.4 Conclusions

In conclusion, we have compiled two datasets containing both computational and experimental reports on magnetic properties of uranium-based binary compounds. Through various data analytics techniques, we have identified several descriptors that are critical for understanding magnetic properties of such systems, even before building any predictive models. These insights were then used in developing families of ML models to predict the magnetic moment size and ordering. We have also extended this approach to other actinides and assessed the performance of the models. Currently the models trained on Dataset I can only predict moment size for AFM ordered structures. Predicting magnetic spin-texture based on the strength of nearest and next nearest neighbor exchange interactions requires additional DFT computations for other magnetic configu-
rations, which is beyond the scope of the current work but nonetheless could be accomplished. Overall, this general prescription, employing both computational and experimental results to construct ML models describing magnetism in actinide-based materials, helps us develop better understanding of structure-property relationships that may exist in such complicated structures.

4.3 Organic Ferroelectrics

4.3.1 Introduction

As already outlined in the previous chapters, ‘soft’ ferroelectric materials, including polymer- and molecular-based crystals are currently becoming viable alternatives to their ceramic counterparts — especially due to the flexibility and low costs of the synthetic chemistry approaches that can be used for their production. It is noteworthy that some of most well known ferroelectric compounds [257], such as Rochelle salt (in which ferroelectricity was originally discovered) and NaNO$_2$ (exhibiting interesting incommensurate structural distortions [258], can actually be classified as molecular crystals. In the discussion accompanying Table 3.3 in Section 3.1.2, we already state that, despite the availability of numerous physical mechanisms that could potentially induce ferroelectricity in polymer and molecular-based systems (in addition to the displacive mechanism which dominates in ceramics), ferroelectricity remains rare — and more of a ‘one-off’ or sometimes
spurious effect in such compounds. Lessons learned by us when studying the DIPA-X and PVDF-based families of electroactive materials provide some understanding as to why this may be so. Even though individual molecular units may be highly polar, there is no guarantee that, when they are joined together into a crystal, the resulting structure will be ferroelectric as (i) large individual dipole moments tend to get canceled and/or (ii) built-in energy barrier for possible molecular reorientation may be way too high to allow ‘easy’ polarization switching among multiple possible directions. In the latter case, although the material may formally be ferroelectric, the associated enormous coercive fields make any practical applications of this effect highly unlikely.

Although specific pieces of knowledge that may be useful for successful cultivation of ferroelectric properties in polymer and molecular-based systems remain elusive — unquestionably, because of the structural richness of the involved materials families and great complexity of the underlying physical mechanisms — we are nonetheless tempted to use the ML frameworks developed and deployed in the previous two Sections to explore this general problem. The main idea is to utilize the already available information of the already known ferroelectric molecular- and polymer-crystals to uncover any common structure – chemical makeup – properties – performance relations and identify possible simple descriptors that would ‘point’ to likely ferroelectric behavior in other molecular- or polymer-based compounds, whether these are fictitious or synthesized already. Strategically, such a study may be a combination of the previous two projects, where any possible connection between the available descriptors and the target functionality (which is some
generic ‘propensity for ferroelectricity’) is totally unclear, plus the associated property of interest (e.g., spontaneous polarization) is a tough, composite material characteristic — i.e., a vector property that is in turn a function of mutual arrangement of local vector moments.

Engaging in this study involves a number of obvious difficulties. We can briefly outline the main two as: (i) Scarcity of the available data on the already known molecular and polymeric ferroelectrics; there are just not too many of them — maybe hundreds, but not thousands, which stipulates the types of ML algorithms that could used — and (ii) Lack of any standards on how such data is organized and curated; i.e., individual research groups report whatever they feel is important in each particular case, rather than some standard set of descriptors, plus different groups can have different quality standards for what is considered to be ‘reliable data.’ On the other hand, in case of such compound crystals that could be readily separated into individual molecule-based structural units, extremely well developed libraries of numerous chemical descriptors could be readily applied to these units, in the same fashion as it was done for the API molecules in Section 4.1. Naturally, chemical descriptors acting on the molecular level can be combined with global, crystal-level descriptors, such as the ones used for the studies of the actinide compounds in Section 4.2. Presumably, the connections that we Humanly expect to make here between the structure and the resulting ‘propensity for ferroelectricity’ may involve the following: (i) Existence of multiple polymorphs for the same compound, some of which should be non-centrosymmetric and polar, while others should be centrosym-
metric and non-polar. (ii) There must exist some ‘mild’ transformation paths connecting
the polar and non-polar phases, meaning that these phases cannot be drastically differ-
ent and should be close enough to each other in energy to allow ferroelectric switching.
Furthermore, while conducting this study, we aim to avoid any heavy calculations —
e.g., like the ones described in Chapter 3 — wherever possible. Although it is clear that
for each new individual ‘prospective ferroelectric’ compound these calculations are ab-
solutely necessary to characterize its properties with full confidence, our goal instead is
to spin out multiple (but not too many!) guesses for such prospective compounds and
provide rough, but computationally cheap estimates of their likely properties.

In the course of this project — which is currently ongoing and will definitely extend
well beyond the scope of this dissertation — we have managed to address some of the
challenges stated above, showcasing how data-driven and ML approaches can be used
to judiciously shortlist molecular- and polymer-based candidates for developing ferro-
electricity from a wide chemical space of different compounds. A number of crystal- and
molecular-level descriptors were found to be important for predicting ferroelectric be-
havior. Furthermore, traditional ML algorithms-based models were used to estimate the
magnitude of spontaneous polarization in the candidate materials without performing
any expensive atomistic simulations. The flowchart including the main developmental
stages of the project is presented in Fig. 4.15.
4. Traditional ML-algorithms based learning

Figure 4.15: Flowchart outlining the main development stages involved in the selection process for identifying new potential molecular- and polymer-based ferroelectrics. Primary stages are shown as grey rectangles, while any necessary secondary stages are represented by white rectangles.
4.3.2 Methods

4.3.2.1 Data Compilation and Curation

The starting point of this study involved compilation of a dataset, called in what follows ‘Dataset I,’ containing records for most molecular- and polymer-based ferroelectrics that have already been reported in the literature [257, 259, 260]. Tables 4.8 and 4.9 list all the collected entries with their corresponding molecular formulas and SMILEs (simplified molecular-input line-entry system), with the latter providing more detailed description of the compound structure and chemical makeup. The Appendix (Chapter 6) contains the diagrams of the chemical structures of these molecular compounds.

At the time of this writing, the compiled dataset contains a total of 76 molecular systems with reported polar properties driven by electronegativity differences between constituting atoms, or presence of secondary interactions such as hydrogen bonding, or extrinsic origins such as a structural instability. The chemical and structural origins of the dataset entries are tremendously diverse and include molecular and polymer crystals, oligomers, co-crystals, clathrates, charge-transfer complexes and derivatives of metal-organic frameworks.

4.3.2.2 Descriptors and ML models

For each of the dataset compounds, we have computed sets of molecular- and crystal-level descriptors. For the former, standard computational MOE [224, 225] 2D chemical
Table 4.8: Dataset listing organic ferroelectrics reported in the literature. Molecular formulas, common names and canonical SMILES are included. This table is continued on the following page.

<table>
<thead>
<tr>
<th>Name</th>
<th>Molecular Formula</th>
<th>Canonical SMILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thioura</td>
<td>CH$_4$N$_2$S</td>
<td>C(=S)(N)N</td>
</tr>
<tr>
<td>Tempo</td>
<td>C$_7$H$_9$NO</td>
<td>C1(C(CCC(N1)[O])(C)C)</td>
</tr>
<tr>
<td>TCAA (Trichloroacetic acid)</td>
<td>C$_3$H$_7$NO$_3$</td>
<td>C1(C=O)N(C=O)N(C=O)N1C(C1)Cl</td>
</tr>
<tr>
<td>Benzil</td>
<td>C$_3$H$_7$O$_2$</td>
<td>C1=CC(C=C1)C=OC(C=O)C2=CC=C2</td>
</tr>
<tr>
<td>DNP (2, 4-dinitrophenol)</td>
<td>C$_6$H$_4$N$_2$O$_3$</td>
<td>C1=CC(C=C1)C=OC(C=O)C2=CC=C2</td>
</tr>
<tr>
<td>TCHM(4-[1-(2-Thienyl)cyclohexyl]morpoline)</td>
<td>C$_7$H$_5$NOS</td>
<td>C1CC(C1)(C2=C(C=CS)N3CCOC=C3</td>
</tr>
<tr>
<td>VDF oligomer</td>
<td>C$_2$H$_4$F$_2$</td>
<td>C(=F)F</td>
</tr>
<tr>
<td>Co-crystal of phenazine-chloranilic acid</td>
<td>C$<em>{12}$H$</em>{10}$N$_2$O$_2$Cl</td>
<td>C1=CC(C=CC1)N=C3C=C3=N2C1=C1(C=OC(C1=O)Cl)OCl</td>
</tr>
<tr>
<td>Co-crystal of phenazine-bromanilic acid</td>
<td>C$<em>{12}$H$</em>{12}$N$_2$Cl</td>
<td>C1=CC(C=OC1)N=C3C=C3=N2C1=C1(C=OC(C1=O)Br)Br</td>
</tr>
<tr>
<td>[H-55DMBP][Hfa]</td>
<td>3C$_2$H$_4$(OH)$_2$.CH$_3$OH</td>
<td>C1=CC(C=CC1)OOCO</td>
</tr>
<tr>
<td>β-quinal-methanol</td>
<td>((CH$<em>2$CF$<em>2$)$</em>{0.65}$(CHF=CF)$</em>{0.35}$</td>
<td>C=C(FC)(C(FC=FC)F</td>
</tr>
<tr>
<td>Nylon-11</td>
<td>C$_1$H$_2$ON</td>
<td>C=C(CCCCNCCCCC)(=O)O</td>
</tr>
<tr>
<td>HdabcoReO$_4$</td>
<td>N$_2$(C$_2$H$_4$)$_2$</td>
<td>C1CNCN1CC2=O[[Re]=O(O)=O</td>
</tr>
<tr>
<td>TGS (Triglycine sulfate)</td>
<td>(NH$_2$CH$_2$COOH)$_2$.H$_2$SO$_4$</td>
<td>C(C(=O)O)(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>TSCC (trisarcosine calcium chloride)</td>
<td>C$_3$H$_5$CaCl$_2$N$_6$</td>
<td>C(CCCCNCCCCC)(=O)O</td>
</tr>
<tr>
<td>Rochelle salt</td>
<td>C$_1$H$_2$KNaO</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>MACA (methyl ammonium chrom alum)</td>
<td>CH$_3$NH$_3$Cr(SO$_4$)$_2$.12H$_2$O</td>
<td>[NH$_4$]+.O.O(O)=O</td>
</tr>
<tr>
<td>MAFeA (methyl ammonium iron alum)</td>
<td>CH$_3$NH$_3$Fe(SO$_4$)$_2$.12H$_2$O</td>
<td>[NH$_4$]+.[O]=O</td>
</tr>
<tr>
<td>MAInA (methyl ammonium indium alum)</td>
<td>CH$_3$NH$_3$In(SO$_4$)$_2$.12H$_2$O</td>
<td>[NH$_4$]+.[O]=O</td>
</tr>
<tr>
<td>Urea chlom</td>
<td>CO(NH$_2$)$_2$.Cr(SO$_4$)$_2$.12H$_2$O</td>
<td>[NH$_4$]+.[O]=O</td>
</tr>
<tr>
<td>GASH (guanidinium aluminum sulphate hexahydrate)</td>
<td>CH$_3$Al(N$_2$O$_5$)$_2$.6H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Guanidinium aluminum selenate hexahydrate</td>
<td>CH$_3$Al(N$_2$S$_2$)$_2$.6H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Guanidinium chromium sulphate hexahydrate</td>
<td>CH$_3$Cr(N$_2$O$_5$)$_2$.6H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Guanidinium gallium selenate hexahydrate</td>
<td>CH$_3$Ga(N$_2$S$_2$)$_2$.6H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Guanidinium gallium selenate hexahydrate</td>
<td>CH$_3$Ga(N$_2$S$_2$)$_2$.6H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Potassium ferrocyanide trihydrate</td>
<td>K$_3$Fe(CN)$_6$.3H$_2$O</td>
<td>[C(=N]C=NC=C(N[=N]C)=N(C)=N(C)=N(C)=N(C)=NC[=N]C=N(C)=N(C)=N(C)=N(C)=N(C)=N(C) ]Fe[=N]</td>
</tr>
<tr>
<td>Potassium ruthenium cyanide trihydrate</td>
<td>C$_6$K$_2$N$_6$Ru$_3$3H$_2$O</td>
<td>[C(=N]C=NC=C(N[=N]C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C) ]Fe[=N]</td>
</tr>
<tr>
<td>Potassium osmium cyanide trihydrate</td>
<td>C$_6$K$_2$N$_6$Os$_3$.3H$_2$O</td>
<td>[C(=N]C=NC=C(N[=N]C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C)=N(C) ]Fe[=N]</td>
</tr>
<tr>
<td>TGS (Triglycine selenate)</td>
<td>(NH$_2$CH$_2$COOH)$_2$.H$_2$Se$_2$O$_4$</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>TGFβ (triglycine fluorobyllate)</td>
<td>(NH$_2$CH$_2$COOH)$_2$.H$_2$Be$_4$F$_4$</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>DGN (diglycine nitrate)</td>
<td>(NH$_2$CH$_2$COOH)$_2$.HNO$_3$</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Glycine silver nitrate</td>
<td>NH$_2$CH$_2$COOHLAgNO$_3$</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Trissarcosine calcium chloride</td>
<td>(CH$_2$NH$_3$CH$_2$COOH)$_3$.CaCl$_2$</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Sodium ammonium tartrate tetrahydrate</td>
<td>NaNH$_2$C$_3$H$_4$O$_6$.4H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
<tr>
<td>Lithium ammonium tartrate monohydrate</td>
<td>LiTIO$_2$H$_2$O$_6$.H$_2$O</td>
<td>C(C(=O)O)(O)=O</td>
</tr>
</tbody>
</table>
Table 4.9: Table 4.8 continued.

<table>
<thead>
<tr>
<th>Name</th>
<th>Molecular Formula</th>
<th>Conventional NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deuterium chloride 2-mercaptoacetate</td>
<td>C₂H₃C(=O)CH₂SH</td>
<td></td>
</tr>
<tr>
<td>Ammonium hydrogen fluoride</td>
<td>NH₄⁺</td>
<td></td>
</tr>
<tr>
<td>Ammonium hydrogen bromide</td>
<td>NH₄⁺</td>
<td></td>
</tr>
<tr>
<td>Ammonium hydroxide iodide</td>
<td>NH₄OH⁻</td>
<td></td>
</tr>
<tr>
<td>Ammonium iodide bromide</td>
<td>NH₄⁺</td>
<td></td>
</tr>
<tr>
<td>Iodide chloride</td>
<td>I⁻</td>
<td></td>
</tr>
<tr>
<td>Bromide chloride</td>
<td>Br⁻</td>
<td></td>
</tr>
<tr>
<td>Chloride chloride</td>
<td>Cl⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate chloride</td>
<td>SO₄²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate monoanion</td>
<td>SO₄¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate di-anion</td>
<td>SO₄²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tri-anion</td>
<td>SO₄³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-anion</td>
<td>SO₄⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-anion</td>
<td>SO₄⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-anion</td>
<td>SO₄⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-anion</td>
<td>SO₄⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-anion</td>
<td>SO₄⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-anion</td>
<td>SO₄⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-anion</td>
<td>SO₄¹⁰⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate undeca-anion</td>
<td>SO₄¹¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate dua-deca-anion</td>
<td>SO₄¹²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tria-deca-anion</td>
<td>SO₄¹³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-deca-anion</td>
<td>SO₄¹⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-deca-anion</td>
<td>SO₄¹⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-deca-anion</td>
<td>SO₄¹⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-deca-anion</td>
<td>SO₄¹⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-deca-anion</td>
<td>SO₄¹⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-deca-anion</td>
<td>SO₄¹⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-deca-anion</td>
<td>SO₄²⁰⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate undeca-deca-anion</td>
<td>SO₄²¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate dua-deca-deca-anion</td>
<td>SO₄²²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tria-deca-deca-anion</td>
<td>SO₄²³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-deca-deca-anion</td>
<td>SO₄²⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-deca-deca-anion</td>
<td>SO₄²⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-deca-deca-anion</td>
<td>SO₄²⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-deca-deca-anion</td>
<td>SO₄²⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-deca-deca-anion</td>
<td>SO₄²⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-deca-deca-anion</td>
<td>SO₄²⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-deca-deca-anion</td>
<td>SO₄³⁰⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate undeca-deca-deca-anion</td>
<td>SO₄³¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate dua-deca-deca-deca-anion</td>
<td>SO₄³²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tria-deca-deca-deca-anion</td>
<td>SO₄³³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-deca-deca-deca-anion</td>
<td>SO₄³⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-deca-deca-deca-anion</td>
<td>SO₄³⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-deca-deca-deca-anion</td>
<td>SO₄³⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-deca-deca-deca-anion</td>
<td>SO₄³⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-deca-deca-deca-anion</td>
<td>SO₄³⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-deca-deca-deca-anion</td>
<td>SO₄³⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁰⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate undeca-deca-deca-deca-anion</td>
<td>SO₄⁴¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate dua-deca-deca-deca-deca-anion</td>
<td>SO₄⁴²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tria-deca-deca-deca-deca-anion</td>
<td>SO₄⁴³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-deca-deca-deca-deca-anion</td>
<td>SO₄⁴⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁰⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate undeca-deca-deca-deca-deca-anion</td>
<td>SO₄⁵¹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate dua-deca-deca-deca-deca-deca-anion</td>
<td>SO₄⁵²⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tria-deca-deca-deca-deca-deca-anion</td>
<td>SO₄⁵³⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate tetra-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁴⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate penta-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁵⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hexa-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁶⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate hepta-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁷⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate octa-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁸⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate nona-deca-deca-deca-deca-anion</td>
<td>SO₄⁵⁹⁻</td>
<td></td>
</tr>
<tr>
<td>Sulfate deca-deca-deca-deca-deca-anion</td>
<td>SO₄⁶⁰⁻</td>
<td></td>
</tr>
</tbody>
</table>
descriptors were used, while descriptors relying on a three dimensional shape of the investigated molecule were excluded due to chances of adding any potential ambiguity to the descriptors space. The list of crystal-level descriptors, as implemented in a Python library [251], included atomic orbital information, such as highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), as well as orbital energies, types of and distances between each pair of atomic sites, fractions of nearest neighbors for each atomic and bond type, statistical data on electronegativity differences between anions and cations, oxidation states of all atoms, valence orbital attributes, such as mean number of electrons in each shell, and orbital field matrix descriptions of the chemical environment of each atom in the unit cell, based on the group numbers, row numbers, distances between coordinating atoms, and Voronoi polyhedra weights. Total energies, polarizations were not calculated for these systems present in Dataset I to avoid expensive computations. Various ML models were built utilizing the computed sets of molecular- and crystal-level descriptors; initially all and consequently only 10 most important descriptors (combined from both sets) were used. The RFR algorithm was employed as implemented in R [91] version 3.4.2 to construct the models. Regression-based algorithms were chosen over classification-based ones, as the endpoint of interest (e.g., magnitude of the spontaneous polarization) is continuous in nature. The choice of RFR was driven by its performance in the previous studies (described in Sections 4.1 and 4.2) when applied to datasets of small size, as well as the algorithm’s capability to shortlist descriptors by importance. That capability is essential for this investigation as it allows
us to identify the likely physical underpinnings enabling ferroelectricity in molecular and polymeric systems of interest.

4.3.2.3 Data Analysis

A histogram plot of the reported spontaneous polarizations exhibited by the compounds present in Dataset I is shown in Fig. 4.16. From this figure, it is evident that it is not only the dataset size that is limited, but also that the dataset is strongly skewed or biased.
towards systems with low values of polarization (which is not completely unexpected, considering usually smaller polarization values displayed by ‘soft’ ferroelectric crystals compared to their ceramic counterparts). Generally speaking, an infusion of many more data entries representing systems possessing high spontaneous polarization is required to remove the bias in the data and, therefore, improve the predictive capabilities of any resulting ML models.

However, the list of existing ferroelectric compounds has already been greatly exhausted in the process of assembling the dataset Therefore, instead we have used an importance sampling approach [261] to generate the extra data as required. This is widely used technique in statistics and other fields for estimating properties of a particular distribution of data. To apply the sampling technique we first found instances for which the polarization magnitude is \( > 5 \mu \text{C/cm}^2 \). The high-dimensional nature of the data poses a challenge in generating additional samples that are similar to these instances. We have used PCA, as detailed in Chapter 2, to reduce the dimensionality of the dataset. The descriptors were transformed into a space where they are mutually independent (different from the original descriptors space). For each of these independent features, we fit a kernel density estimate (with gaussian kernels) to approximate their distributions. We draw 100 new sample points from these distributions which are then subsequently inverse transformed to the original high-dimensional descriptors space. Note that these generated fictitious samples are used only to train the ML models, while the test set still consists of the unaltered original data collected for the actual materials. A histogram plot
4. Traditional ML-algorithms based learning

Figure 4.17: Histogram partitioning of Dataset II, including both the original (blue bars) and newly generated entries with importance sampling (orange bars), by magnitude of spontaneous polarization.

including the extra generated data along with the original data is presented in Fig. 4.17, showing that the combined dataset, called in what follows ‘Dataset II,’ is now less biased towards structures with low polarization magnitudes.
4. Traditional ML-algorithms based learning

4.3.3 Results and Discussion

The first set of ML models was built using Dataset I where one model was constructed with molecular descriptors only and another with crystal-level descriptors. The average RMSEs of these ML models for predicting spontaneous polarization are on the order of 3.9–4.1 μC/cm². The average RMSE of the best ML model that was constructed utilizing
10 most important descriptors taken from the combined set of molecular- and crystal-level descriptors is 3.30 $\mu$C/cm$^2$. The learning curve for this model is shown in Fig. 4.18, while its predictive accuracy estimates are compiled in Table 4.10.

To improve the predictive accuracy of the ML models described in the previous paragraph, we constructed another set of models employing Dataset II. Table 4.10 lists average RMSE errors for these models as well. It is evident that utilization of the importance sampling approach to un-bias Dataset I has improved the predictive accuracy of the models, reducing the average test set RMSE error to 1.84 $\mu$C/cm$^2$. The combined list of most important descriptors, as identified by RFR includes information on ionization potentials estimated using Gearing autocorrelations [262], hydrogen bonding interactions and electronegativity differences.

We note that it is possible to create different variants of Dataset II by changing the cut-off value of polarization between 5 $\mu$C/cm$^2$ and $\sim$20 $\mu$C/cm$^2$. However, due to much smaller number of reported structures with polarization of $>10$ $\mu$C/cm$^2$, the generated samples in datasets with large cut-offs will be too similar in nature and therefore will not

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$E_{\text{Test}}(30)$</th>
<th>$E_{\text{Test}} - E_{\text{Train}}(30)$</th>
<th>$E_{\text{Test}}(N_{\text{max}})$</th>
<th>$E_{\text{Test}} - E_{\text{Train}}(N_{\text{max}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset I</td>
<td>5.01</td>
<td>2.84</td>
<td>3.30</td>
<td>0.93</td>
</tr>
<tr>
<td>Dataset II</td>
<td>4.13</td>
<td>1.76</td>
<td>1.84</td>
<td>0.32</td>
</tr>
</tbody>
</table>
add any substantial information to the dataset. Consequently, the subsequent ML models developed using such datasets will not significantly improve their predictive accuracy. A different strategy may involve developing ML models based on datasets, including only compounds sharing the same mechanisms for developing ferroelectricity. This approach can further improve prediction accuracy, but at present our capabilities for creating such models is limited by the size of the original dataset, which will be reduced to subsets of less than 20 entries in size when partitioned by the ferroelectricity mechanism type.

4.3.3.1 Shortlisting potential organic ferroelectrics

In addition to predicting the magnitude of the spontaneous polarization in molecular- and polymer-based compounds, this project aims to develop strategies for shortlist potential candidates that may exhibit ferroelectricity by identifying important descriptors related to their structural and electronic properties. The following criteria were used to shortlist the most attractive candidates from the list of 1,034,175 organic compounds present in the CSD [52].

In order to qualify to be shortlisted as a potential ferroelectric, a compound must have at least two different structural phases belonging to centrosymmetric and polar space groups, respectively. The list of crystallographic polar point groups includes 10 groups: 1(C1), 2(C2), 4mm(C4v), 3(C3), 4(C4), 3m(C3v), 6(C6), m(C5), mm2(C2v) and 6mmm(C6v). The centrosymmetric point groups are 1, 2/m, mmm, 4/m, 3, 3m, 6/m, 6/mm, m3m and m3m. The presence of both centrosymmetric and polar phases for the same material is similar to
the cases of ceramic ferroelectrics, such as BTO and PTO, where (displacive) polarization develops through a continuous phase transformation between the two phases.

Utilizing the symmetry criterion for having both centrosymmetric and polar polymorphs, we identified a total of 8,997 compounds. To further reduce the number of compounds, we have used the following constraints to this new list.

1. Compounds with heavy elements, such as Mo, As, Au, Ba, Dy, Eu, Ga, Gd, Ge, Hg, La, Nd, Pb, Rh, Ru, Sb, Se, U, Ir, Os, Pd, Pt, Re, Ta, Sm, Tb, Te, U, W, Yb, Y, Sb, Tc, Tm, Co, Nb, In, V and Be, were discarded.

2. Toxic elements, such as Cd, Pb, Th, Hg and Sb, were excluded.

3. Hydrates and charged complexes were discarded.

4. (This step may be skipped in the future to potentially include more interesting compounds into the mix) Only compounds containing atoms belonging to exclusively to the first row of the periodic table were kept in the list.

A total of 4,932 compounds remained in the final list, for which the spontaneous polarization could be quickly predicted utilizing the previously developed ML models.

The step involving the prediction of the system polarization requires access to structure description for each of the compound phases. At present, we do not have full access to such information in the CSD. However, to demonstrate the predictive capabilities of the developed ML models, we have created a further reduced list of compounds that
Table 4.11: Potential organic ferroelectrics for which only 2 structural phases are present in the CSD. Molecular formulas, canonical SMILEs and predicted spontaneous polarization values (in $\mu\text{C/cm}^2$) are provided.

<p>| Molecular Formula | Canonical SMILE | $|P_s|$ |
|-------------------|-----------------|------|
| $\text{C}<em>{10}\text{H}</em>{15}\text{N}<em>{5}\text{O}$ | $\text{CC(\text{CNC1=NC=NC2=C1NC=\text{N}2})CO}$ | 10.39 |
| $\text{C}</em>{10}\text{H}<em>{6}\text{O}</em>{8}$ | $\text{Cl=C(=C=\text{C1C}(=\text{O})\text{O})C(=\text{O})C(=\text{O})C(=\text{O})C(=\text{O})C(=\text{O})}$ | 10.41 |
| $\text{C}<em>{13}\text{H}</em>{21}\text{NO}<em>{8}$ | $\text{CCOC1=C(OC(C1\text{N}C(=\text{O})C(C(C\text{O})\text{O})C(=\text{O})}$ | 10.09 |
| $\text{C}</em>{13}\text{H}<em>{21}\text{N}</em>{5}$ | $\text{ClCCC(C1)C2=CC(=\text{NC}(=\text{N2})\text{N3CC}(\text{C}3)\text{N}}$ | 10.14 |
| $\text{C}<em>{13}\text{H}</em>{21}\text{N}<em>{3}\text{O}$ | $\text{Cl=CC(=CN=C1)OC2=CC(=C(C=C2)\text{C#N})C#N}$ | 11.17 |
| $\text{C}</em>{15}\text{H}<em>{10}\text{N}</em>{2}\text{O}<em>{6}$ | $\text{COC(=\text{O})C1=C2C(=\text{NC3=C(N2C}(=\text{O})C=C(C3=\text{CO})C(=\text{O})C=C1}$ | 11.14 |
| $\text{C}</em>{16}\text{H}<em>{14}\text{F}</em>{3}\text{NO}<em>{2}$ | $\text{ClCC2=CC(=\text{C4=C2N(C1)CCC4=CC=C(C4)\text{C}#N})C#N}$ | 11.27 |
| $\text{C}</em>{17}\text{H}<em>{19}\text{N}</em>{3}\text{O}<em>{4}$ | $\text{Cl(COCI}1\text{NC2=NC3=CN}=\text{C2N=C(N=C30CC4=CC=CC=C4)NOC})O$ | 12.65 |
| $\text{C}</em>{18}\text{H}<em>{20}\text{N}</em>{4}\text{O}<em>{3}$ | $\text{CC1(N=CN=C1)CC}1\text{C2=CC(=C2OCC3=CC(=C3)\text{N+}=(O-))N})\text{N}$ | 10.40 |
| $\text{C}</em>{19}\text{H}<em>{11}\text{NO}</em>{3}$ | $\text{Cl=CC(=C(C=Cl)=N30CC4=CC=C(C=NC=C3N=CC=C4)\text{C}#N})C#N}$ | 12.76 |
| $\text{C}<em>{19}\text{H}</em>{12}\text{BF}<em>{3}\text{N}</em>{3}\text{O}$ | $\text{B(C1=C(N=C(C1)\text{F})C2=C(N=C(C2)\text{F})OCC3=CC=C(C=3N=CC=C4)}=\text{C}=\text{C4}$ | 10.99 |
| $\text{C}<em>{19}\text{H}</em>{22}\text{N}<em>{3}\text{O}</em>{5}$ | $\text{CC1=NOC(=C1)CCCCCCCO2=CC=C(C=C2)C3=NC(C03)\text{CO}}$ | 11.11 |
| $\text{C}<em>{19}\text{H}</em>{30}\text{O}<em>{7}$ | $\text{CC}(\text{C}(\text{OC})\text{OC}1\text{C}=\text{C2C}(=\text{C1})\text{C}(=\text{C2})\text{OC2}=\text{OC}(\text{C}00)\text{COCO})$ | 11.01 |
| $\text{C}</em>{22}\text{H}<em>{24}\text{N}</em>{3}\text{O}<em>{3}$ | $\text{C}(=\text{O})(\text{C}(=\text{O})\text{O})\text{N}$ | 11.18 |
| $\text{C}</em>{22}\text{H}<em>{28}\text{NO}</em>{3}$ | $\text{C}(=\text{O})\text{N}$ | 11.23 |
| $\text{C}<em>{22}\text{H}</em>{22}\text{O}<em>{2}\text{H}</em>{2}$ | $\text{CC(=O)C1CC(=O)=C}=\text{CC}(=\text{O})\text{OC(=C}1\text{CC}(=\text{O})\text{C}=\text{CC}(=\text{O})\text{O}1)\text{CC}(=\text{O})\text{C}$ | 12.79 |
| $\text{C}<em>{23}\text{H}</em>{27}\text{F}<em>{3}$ | $\text{C}(\text{CCC}(\text{F})(\text{F})\text{OC}=\text{CC=C(CC=C(=\text{O})\text{O})}$ | 10.15 |
| $\text{C}</em>{23}\text{H}<em>{20}$ | $\text{CCCCCC}(=\text{C}1\text{CC}=\text{C(C}1\text{CC}=\text{COC}(=\text{O})\text{O})\text{O})\text{O}$ | 11.01 |
| $\text{C}</em>{22}\text{H}<em>{22}\text{N}</em>{3}\text{O}<em>{4}$ | $\text{C1=CC=C2C(=\text{Cl})C(=\text{O})\text{OC}(=\text{N}2)\text{C}=\text{CC}=\text{C=C}=\text{C}3\text{N}=\text{CN}=\text{C5}\text{C}(=\text{O})\text{O}4$ | 9.85 |
| $\text{C}</em>{23}\text{H}<em>{25}\text{F}</em>{3}\text{N}<em>{2}\text{O}</em>{3}$ | $\text{ClCN(=CC1\text{C}=\text{C=C}(=\text{O})\text{F})3C\text{CN}(=\text{O})\text{C4}=\text{CC}=\text{C(C}4\text{)}\text{OC(F)}$ | 11.20 |
| $\text{C}<em>{24}\text{H}</em>{28}\text{N}<em>{6}\text{O}</em>{2}$ | $\text{CCCO1=CN}=\text{CC}(=\text{C1})\text{C}=\text{NN}=\text{C3N}2\text{C}=\text{C}(=\text{C}4)\text{CN}5\text{CCOC5}N=\text{C3C}$ | 10.74 |
| $\text{C}<em>{24}\text{H}</em>{34}\text{O}_{15}$ | $\text{CC}(=\text{O})\text{OCC1C(C(C}01\text{)C2(=\text{CCOC2})OC}(=\text{O})\text{OC}(=\text{O})\text{O}1\text{OC}(=\text{O})\text{OC}1\text{OC}(=\text{O})\text{OC}(=\text{O})$ | 10.68 |</p>
<table>
<thead>
<tr>
<th>Molecular Formula</th>
<th>Canonical SMILE</th>
<th>$P_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C$<em>{24}$H$</em>{35}$N$_3$</td>
<td>CC1=C2CCCCN2C(=C1)C3=CC=C(C=C3)N(C)CCCCN4CCCCC4</td>
<td>10.66</td>
</tr>
<tr>
<td>C$<em>{25}$H$</em>{32}$N$_2$O$_2$</td>
<td>CC(C)N1CC(C(C1)NC(=O)C2=CC=CC=C3N2CC4=CC=C(C=C4)OC</td>
<td>9.94</td>
</tr>
<tr>
<td>C$<em>{26}$H$</em>{32}$NO$_3$</td>
<td>$[B$-$][N+]$=C(C=CC=C(C(=C1)C=C(C=C1)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>9.51</td>
</tr>
<tr>
<td>C$<em>{27}$H$</em>{39}$</td>
<td>C12=C3C(=CC=C2)C4=CC=C5=C4(=CC=C5)C3=CC1=C7=CC=C7=C6</td>
<td>9.45</td>
</tr>
<tr>
<td>C$<em>{28}$H$</em>{33}$N$_2$O$_2$</td>
<td>C1=CC=C(C=CC=C(C=C2)C3=CC=C(C=C3)C4=CC=C(C=C4)OC</td>
<td>10.82</td>
</tr>
<tr>
<td>C$<em>{29}$H$</em>{16}$N$_3$O</td>
<td>C12=CC=C(C=CC=C(C=C1)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.37</td>
</tr>
<tr>
<td>C$<em>{30}$H$</em>{28}$O$_5$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.20</td>
</tr>
<tr>
<td>C$<em>{31}$H$</em>{26}$N$_3$O</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.46</td>
</tr>
<tr>
<td>C$<em>{32}$H$</em>{24}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.18</td>
</tr>
<tr>
<td>C$<em>{33}$H$</em>{33}$N$_3$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>10.75</td>
</tr>
<tr>
<td>C$<em>{34}$H$</em>{22}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.09</td>
</tr>
<tr>
<td>C$<em>{35}$H$</em>{34}$N$_3$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.45</td>
</tr>
<tr>
<td>C$<em>{36}$H$</em>{28}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>10.95</td>
</tr>
<tr>
<td>C$<em>{37}$H$</em>{31}$N$_3$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>10.93</td>
</tr>
<tr>
<td>C$<em>{38}$H$</em>{29}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>11.17</td>
</tr>
<tr>
<td>C$<em>{39}$H$</em>{27}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>9.41</td>
</tr>
<tr>
<td>C$<em>{40}$H$</em>{25}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>9.17</td>
</tr>
<tr>
<td>C$<em>{41}$H$</em>{23}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>10.83</td>
</tr>
<tr>
<td>C$<em>{42}$H$</em>{19}$O$_2$</td>
<td>C12(CC3=C(C=CC=C3C(=O)C=C2CC(=O)C(C=CC=C3)C5=C(CC=C(C=C5)C6=CC=C(C=C6)OC</td>
<td>10.93</td>
</tr>
</tbody>
</table>
Figure 4.19: Histogram showing predicted polarizations of compounds from Tables 4.11 and 4.12. Polarizations for DIPA-Cl and 100% crystalline β-PVDF are shown for comparison as well. Note that typical polarizations of (partially amorphous) poled experimental samples of β-PVDF or PVDF/TrFE are at least twice as weak.

possess only two structural phases (one centrosymmetric and the other polar) with the same lattice type. This list contains 45 individual compounds listed in Tables 4.11 & 4.12. The [centrosymmetric, polar] space groups that these structures crystallize in are: [C2/c, C221], [P62c, P2/c], [C2/c, Cc], [P2/c, P21], [C2/c, Cmc21], [P3m1, P21], [C2, C2/c], [C2/c, Fdd2], [C2, C2/c], [P2/c, P21], [C2/c, Cmc21], [C2/c, Cc], [P2/c, P21], [C2, Cmca], [C2, C2/c], [C2/c, Cmc21], [C2/c, C221], [C2/c, Cc], [C2, C2/c], [C2, C2/c], [C2/c, Cc],
For this short list of compounds, we were able to generate the canonical SMILEs and molecular formulas using PubChem [263] and OpenBabel [264, 265] libraries, which then allowed us to estimate the magnitudes of their spontaneous polarizations. The average RMSE for these predictions is \( \sim 3.65 \mu \text{C/cm}^2 \) and the magnitude varies between 9.2 and 12.8 \( \mu \text{C/cm}^2 \). A histogram plot of predicted polarizations for these 45 compounds is shown in Fig. 4.19. We note that the quality of these predictions may vary depending on whether the compound structures are obtained directly from CSD or created using PubChem and OpenBabel, with the former being a more accurate approach. At this point, the list of prospective ferroelectric compounds could be reduced even further, e.g., by computing the energy differences between the polar and non-polar crystal configurations, and removing any entries for which these difference are too high (and thus presumably indicate that the polarization in the corresponding structures would be difficult or impossible to switch). However, since such a step actually involves performing DFT-based computations and thus would be rather computationally expensive, we have not performed it in this investigation. On the other hand, such computations are standard for DFT and could be easily carried out when necessary.
4.3.4 Summary

In conclusion, we have compiled datasets containing reports on molecular and polymer-based ferroelectric compounds from the available literature. Through various data analytics techniques, we have identified several descriptors that are critical for understanding polar properties of such systems and developed families of ML models that could predict the magnitude of spontaneous polarization in them. We have also applied data mining to shortlist potential organic ferroelectrics based on their polymorphism, symmetry and chemical makeup, and predicted the values of spontaneous polarization for them. Further investigations including checking the stability of the ground-state configurations and height of the energy barriers between structural phases can be easily accomplished and will significantly enhance the presented analysis, but are beyond the scope of this project as it is currently envisioned. Overall, the developed general prescription, employing both data-driven sampling techniques and ML models describing polar properties of organic crystals, helps us improve our understanding of mechanisms governing the emergence of such complex material functionalities, as well as identify potential candidates possessing such functionalities for further in-depth investigations.

4.4 Chapter Summary

The physical insights obtained from the Human Learning studies have facilitated three different investigations to predict advanced functionalities of materials systems — molec-
ular crystallization propensity, magnetic moment strength and arrangements, and predisposition for ferroelectricity. Each of these functionalities relates to a variety of applications, including those in pharmaceutical sciences or modern electronic devices. We have employed several traditional ML algorithms to construct predictive models. Crystallization propensities of small organic molecules were predicted utilizing appropriate data-driven techniques. We have identified a set of molecular descriptors and experimental factors that are important for understanding the inherent crystallization tendencies of APIs. While the ML models built solely on API related data can predict the endpoint with 30% RMSE, inclusion of solvent information into such models has improved this accuracy to 20% RMSE. Beyond the insertion of solvents, the presence of impurities and/or degradants has the greatest influence on the model accuracy.

In our next investigation, the origins of magnetic behavior in actinide systems were studied on the electronic level using datasets comprising both computational and experimental data. The structure-property links for the families of uranium- and neptunium-based binary compounds were established. Magnetic moment size (spin and orbit) was predicted with ~85% accuracy, whereas the models to predict general magnetic ordering had accuracy of 76%.

In the last case study, we developed a framework utilizing a combination of data-mining and ML-based algorithms to predict the magnitude of spontaneous polarization in molecular- and polymer-based compounds. Collation and analysis of available reports (both computational and experimental) on existing ‘soft’ ferroelectric materials
were essential for this study. During the ML model construction, a specific combination of molecular- and crystal-level descriptors was identified as most important for predicting ferroelectric polarization, which could be accomplished with an average accuracy of $\sim 80\%$. The developed model was then utilized to conduct a (computationally inexpensive) screening of a large variety of molecular crystals to shortlist materials that may potentially exhibit ferroelectricity.

In conclusion, it is evident that general analytic frameworks combining ML algorithms with existing data can be used to predictively evaluate a wide range of properties for diverse families of materials. In order to make reliable predictions about materials behavior, ML-based learning methods require that appropriate data be assembled and curated (e.g., evaluated and ranked for quality), and that sensible descriptors, representing the underlying physical phenomena underpinning the properties in question, be established. Here, a unified ML-based modeling framework was developed and applied to three cases emphasizing different target functionalities in highly dissimilar materials families. Specifically, we predicted (i) crystallization propensity of small organic molecules, (ii) strength and ordering of magnetic moments in actinide-based compounds and (iii) physical origins behind the emergence of ferroelectricity in molecular and polymeric crystals. Despite exploring profoundly different (and difficult to predict with the help of intuitive strategies) targets, all case studies adopted a uniform approach to data curation, selection and design of descriptors, and construction of ML models. The produced results are expected to significantly contribute to both applied (better understanding of materials properties)
and fundamental (ML algorithms) research in materials science.
Chapter 5

Lessons learned and concluding remarks

After an examination of the projects presented in this dissertation, it is evident that data-driven approaches, working in combination with conventional investigative methods in the fields of physics and materials science, could be useful tools for enhancing our understanding of materials behavior. On the other hand, just as for any tool, such approaches must be competently applied and their limitations should be kept in mind and respected — in order to achieve meaningful results. In this regard, careful data curation appears to be essential. It also often a challenging step towards increasing predictive accuracy, since the utilization of inaccurate, unreliable or biased input data in the process of ML model construction and validation can produce equally hazy outcomes that are difficult to understand and interpret.

While experimental reports could serve as reasonable sources for initial datasets and further workflow construction, they are usually highly heterogeneous in content and
quality. That is, they may lack specific information needed, while for the available data it may be nontrivial to assign confidence levels (or decide whether or not and to what degree to trust the data). More importantly, in many instances, such as demonstrated in the case of molecular- and polymer-based ferroelectrics, datasets built on experimental observations may not even have enough data points after curation to construct reliable predictive models. On the other hand, datasets constructed with the help of computer simulations and modeling (such as, e.g., DFT-based atomistic techniques) could be highly homogeneous in content and quality, as well as include data that is difficult to obtain experimentally. Therefore, it is beneficial to have a combination of both experimental and computational data when developing and creating generalized workflows for exploring material properties and accelerating materials design and discovery.

The other important issues (besides data curation) in utilizing the ML tools efficiently involve the choice of appropriate data-processing algorithms and model interpretability. As already mentioned, for very large datasets, deep learning strategies are generally preferred for exploring any possible trends present in the data. However, interpretability of such models could be quite low, as the established connections within the data may often be too complex. Therefore, the relationships between material structure / composition and properties / behavior produced by the deep learning models may lack any apparent physical meaning and thus provide no useful insights.

Utilization of traditional ML algorithms can then be the most reasonable approach for building easy to interpret (at least, as we would hope) predictive ML models, when the
available datasets are small to moderate size, as is quite common in the field of materials science. In this situation, identification of most important relations connecting input and output of the models becomes more straightforward and could provide relevant physical insights into the nature of explored materials.

Finally, performance optimization of the ML models could also be a significant collection of issues, including evaluation and improvement of bias and variance present in the data, as well as a variety of ‘shortlisting’ tactics allowing one to avoid lengthy calculations when necessary.
Chapter 6

Appendix

6.1 Specifications of ML Models

6.1.1 Hyperparameters used for ML models constructed in Chapter 4.1

Randomforest package as implemented in R version 3.4.2[91] was used to build all models using RFR algorithm. The number of trees was varied from 20 to 100, with the number of cross validations ranging from 2 to 10. Optimal prediction accuracy was obtained for combinations involving 40 decision trees with 5 cross validations. The split point used was 5 with maximum depth of 20. Bootstrap aggregation with replacement was used for sampling the data and the decision tree classifier used was 1 (default). Another package known as e1071[91] was used to build models using SVMR algorithm in the same version of R. A standard grid search was followed to find optimized values.
for the two tunable parameters such as the cost (500) and epsilon parameters (0.05). The package Neuralnet[91] as implemented in R that uses a feed-forward neural network with 2 hidden layers with 100 repetitions was applied for building the models with NN algorithm.

6.1.2 Hyperparameters used for ML models constructed in Chapter 4.2

The optimized hyperparameters as obtained using a grid-search method for the regression algorithms are listed below. The amount of penalization ($\alpha$) is 0.01 for LASSO. KRR uses a linear kernel with a regularization constant of 0.1 SVMR based models has used cost and epsilon parameters of 500 and 0.04 respectively. RFR has used 60 decision trees and all the algorithms are also subjected to 10 cross-validations for each model. RFC has used 40 decision trees and all the algorithms are also subjected to 10 cross-validations for each model.

6.1.3 Structures of molecular compounds
Figure 6.1: 3D structures of molecular compounds present in Dataset I as detailed in Chapter 4.3.
Figure 6.2: 3D structures of molecular compounds present in Dataset I as detailed in Chapter 4.3.
Figure 6.3: 3D structures of molecular compounds present in Dataset I as detailed in Chapter 4.3.
Figure 6.4: 3D structures of molecular compounds present in Dataset I as detailed in Chapter 4.3.
References


6. Appendix


6. Appendix


185. Rana, D. S. *et al.* Implications of phase-segregation on structure, terahertz emission and magnetization of Bi(Fe$_{1-x}$Mn$_x$)O$_3$($0 < x < 0.5$) thin films. *EPL* **84**, 67016 (2009).

186. Choi, E.-M. *et al.* Strong room temperature magnetism in highly resistive strained thin films of BiFe$_{0.5}$Mn$_{0.5}$O$_3$. *Appl. Phys. Lett.* **98**, 012509 (2011).


205. Khomchenko, V. *et al.* Synthesis and multiferroic properties of Bi$_{0.8}$A$_{0.2}$FeO$_3$ (A = Ca, Sr, Pb) ceramics. *Appl. Phys. Lett.* **90**, 242901 (2007).


