MACHINE LEARNING IN MATERIAL SCIENCE

By

AJAY GOWRIPURA VIJAYAKUMAR

A DISSERTATION SUBMITTED TO

THE UNIVERSITY OF LIVERPOOL

in partial fulfillment of the requirements
for the degree of

MASTER IN SCIENCE
DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

22/11/2024

STUDENT DECLARATION

I hereby certify that this dissertation constitutes my product, that where the language of others is set forth, quotation marks so indicate. That appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I confirm that I have not copied material from another source, committed plagiarism, commissioned all or part of the work (including unacceptable proofreading), or fabricated, falsified, or embellished data when completing the attached piece of work.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Ajay Gowripura Vijayakumar

ACKNOWLEDGEMENT

The completion of this dissertation would not have been possible without the guidance, support, and encouragement of several key individuals. I take this opportunity to express my sincere gratitude to those who have contributed significantly to this milestone in my academic journey.

First and foremost, I owe immense gratitude to my supervisor, **Prof. Vladimir Gusev**, for his exceptional guidance and patience throughout this research. His constructive feedback and expertise have been instrumental in shaping the direction of this dissertation. I am privileged to have had the opportunity to learn from his knowledge and experience, which have enriched my understanding and approach to research.

I am deeply indebted to my parents, **Miss Rudramma J. and Mr. G.T. Vijayakumar**, for their unwavering support, encouragement, and sacrifices. Your belief in my abilities has been the cornerstone of my achievements. I am profoundly grateful for your love, guidance, and inspiration, which have motivated me to pursue my goals with determination.

I extend my heartfelt thanks to my guardians, **Miss Shobha G. and Prof. B. J. Gireesha**, for their invaluable mentorship and encouragement. Your wisdom and guidance have been a source of strength throughout my journey. Your insights and advice have greatly influenced my academic and personal growth. Your unwavering belief in my potential has been a constant source of motivation.

My appreciation also extends to the esteemed **Faculty Members of Data Science and Artificial Intelligence** program at the University of Liverpool. Their teaching and mentorship have provided me with the knowledge and skills essential to undertaking this research. Their dedication to academic excellence has been a source of inspiration and has greatly contributed to my learning experience.

A special heartful thanks to **Shalini Palakshaiah, Durgambha H. M., and Deekshitha J**. for their unwavering support, encouragement, and belief in me throughout my academic journey. Their thoughtful guidance and constant reassurance have been invaluable during challenging times. I am truly fortunate to have had their support and friendship, which have been a source of strength and positivity.

This dissertation represents the culmination of collective efforts and support. To all those who have contributed, I offer my deepest gratitude. Your encouragement and belief in my abilities have made this achievement possible, and I remain profoundly thankful for your role in my journey.

Machine Learning in Material Science

ABSTRACT

This dissertation investigates the application of machine learning, specifically a Random Forest Regressor model, in predicting key properties of perovskite materials, which are highly promising for renewable energy applications due to their superior light-absorption characteristics. The project's objective is to use machine learning to accelerate material discovery by predicting essential material properties like stability, efficiency, and optoelectronic qualities based on compositional and structural data. The study primarily applies featurization techniques such as site-stats fingerprinting and X-ray Diffraction calculator to extract the relevant features, giving a robust dataset enabling the model to find patterns and optimize perovskite characteristics.

The Random Forest Regressor was chosen for its interpretability and efficiency with non-linear datasets typical of materials science. The model is trained and validated on MatBench, a database with comprehensive materials data, which supports the model's ability to generalize across various perovskite compositions. Combining the featurization methods with Random Forest's ensemble approach, the study achieved strong predictive performance, accurately forecasting material properties critical to perovskite performance in solar cells.

The results show that machine learning integration can speed up the prediction of material properties, effectively reducing traditional research and testing timelines. This approach represents an advance in AI for materials science, providing a route to harness the predictive insights that will contribute to more sustainable material design. These findings add significantly to renewable energy research, supporting the creation of environmentally friendly technologies critical to Global Sustainability Objectives.

STATEMENT OF ETHICAL COMPLIANCE

The research in this dissertation follows the ethical policies and standards of the institution regarding handling and analyzing data. The project uses material data from public sources, namely the MatBench database, classified as Ethical Compliance code A0, meaning no human data involvement that would pose an ethical risk or require participants. No human subjects were involved, nor is any sensitive or identifiable information disclosed in this research study.

All data processing and analyses, and model developments were performed in consideration of ethics in the best interest of responsible practices for research to ensure data integrity, transparency, and reproducibility. The research fully adheres to institutional guidelines for the responsible use of scientific data in an ethical manner.

Contents

Table of Figures

1. INTRODUCTION

Materials science is a key enabler of technology development, especially when it comes to renewable energy. With the need for transitioning toward clean energy, efforts in materials science will increasingly shift toward developing the building blocks of energy devices, such as solar cells. Among several emerging material classes, perovskites have gained huge interest in the research domain of photovoltaics. With their typical ABX_3 crystal configuration, perovskite materials boast excellent light-absorbing qualities, ease in fabrication, and relatively low production costs compared to conventional materials like silicon. In the last few years, there has been a rapid improvement in perovskite solar cell efficiency; thus, it can well be considered as a nominee for the next generation of photovoltaic technologies.

Despite their promise, perovskite materials still face significant challenges that prevent their widespread commercial application. These are mainly centered around the stability and long-term performance at the device level. Environmental factors, such as humidity, heat, and light exposure, easily degrade perovskites and seriously lower their performance over time. In the process of addressing such challenges, there is an increasing demand for techniques that, preferably before actual laboratory synthesis, can predict the properties of perovskite materials. This will not only speed up the discovery of new materials but also increase the efficiency and stability of the perovskite-based solar cells as a whole.

Traditionally, the discovery of materials has been mainly empirical, with new materials being synthesized and then tested in the laboratory. Such processes are slow to realize, consume many resources, and are very expensive. This trend is increasingly unsustainable, as more materials will be required to meet the global challenges of today, including those linked to climate change. In the last few years, techniques using machine learning have emerged as strong tools to speed up material discovery. ML models can analyze big datasets of material properties and predict the performance of new materials by their composition and structure. This approach can potentially reduce the time and cost associated with the discovery of new material properties dramatically, especially when complemented with computational methods that can generate big datasets of material properties.

The focus of this research is to apply machine learning to the prediction and optimization of perovskite materials for solar cell applications. More precisely, the work is targeted at the

development of predictive models that could forecast key properties such as efficiency, stability, and optoelectronic characteristics of perovskite materials, using their chemical composition and crystal structure. Patterns and relationships in large datasets of properties of perovskites have to be learned by machine learning models to guide the search for more efficient and stable perovskite materials.

1.1 Problem Statement

While perovskite materials indeed have shown very impressive performance so far, there is still a lot that is not well understood in the interrelationships of composition, structure, and properties of these materials. Even minor variations in chemical composition or crystal structure can make great differences in performance. This makes designing new materials that will exhibit the desired properties for specific applications, such as high efficiency and stability in solar cells, challenging. Testing of materials experimentally for new ones is time-consuming and thus expensive; hence, the process of identifying the best candidates for further development cannot be quickly attained.

Machine learning has provided a promising solution to the challenges by enabling one to develop predictive models that can estimate material properties based on structural and compositional data. But for that, machine learning models have to be fed with meaningful features representing the structure and composition of the material, from large datasets, for the former to become viable predictive models of perovskite material properties. Such meaningful features can be captured from more sophisticated feature extraction methods such as site-stats-fingerprint and X-ray diffraction (XRD) calculators.

This work seeks to overcome these challenges through the use of machine learning methods to study the prediction of properties in perovskite materials. It will particularly use the RFR algorithm in developing a predictive model capable of forecasting the performance of perovskite materials by leveraging feature extraction methods that capture key structural and compositional information. This study will develop a model that can accurately predict material properties and guide the design of new, more efficient perovskite materials.

1.2 Motivation

The purpose of this work is to accelerate the creation of a class of perovskite materials for the use of renewable energy, especially solar energy. At the pace at which the globe is requiring one to transit into sustainable energy sources, the materials that can facilitate these solar cells efficiently and cost-effectively are in high demand. Perovskites have shown great promise in this respect, but their development into new materials with high efficiency and good long-term stability remains one of the major challenges.

Enabled by machine learning techniques, this work aims to accelerate material discovery by reducing tedious experimental testing. Machine learning models can help in advance to predict the properties of new, hitherto synthesized materials and guide the researcher in selecting the most promising candidates for further experimental investigation. Such models can also unveil unexplored relationships between the composition and structure of a material and its performance, thus providing new avenues toward the optimization of perovskite materials for applications in solar cells.

Ultimately, it is hoped that the research will contribute to developing more efficient and stable perovskite materials and thus pave the way to further the field of renewable energy and global sustainability goals. This work applies machine learning to the prediction of material properties and thus aims at valuable insight accelerating the design and optimization of perovskite-based solar cells and other optoelectronic devices.

1.3 Aims and Objectives

The main aims and objectives of this research are as follows:

- **Develop Predictive Models for Perovskite Materials**: To create machine learning models that can predict key properties of perovskite materials, such as efficiency, stability, and optoelectronic characteristics, based on their chemical composition and crystal structure.
- **Leverage Feature Extraction Techniques**: To employ advanced feature extraction methods like site-stats-fingerprint and X-ray diffraction (XRD) calculators to capture structural and compositional information from large datasets of perovskite materials.

- **Utilize Random Forest Regressor for Prediction**: To apply the Random Forest Regressor (RFR) algorithm, a powerful machine learning tool, to develop an accurate predictive model for perovskite material properties.
- **Optimize Model Performance**: To refine the predictive models through data preprocessing, hyperparameter tuning, and performance evaluation to ensure the highest level of accuracy.
- **Contribute to Sustainable Material Design**: To provide insights that can guide the design of more efficient, stable, and cost-effective perovskite materials for solar cells and other renewable energy applications.

2. BACKGROUND READING

2.1 Perovskite Materials

Perovskites belong to the group of materials that share the same crystal structure, ABX_3, with A as the large cation-often an alkali or alkaline earth metal. B being a smaller cation, typically a transition metal and X as an anion, with the most common members being halides, such as iodine, bromine, or chlorine. This class of materials is very exciting due to their versatility and tunability for numerous applications in renewable energy and optoelectronics. (Kojima, A., Teshima, K., Shirai, Y., & Miyasaka, T. (2009)).

The most famous perovskite material in modern research is lead halide perovskite-e.g., $CH_3NH_3PbI_3$- which has exhibited exceptional efficiency as the light-absorbing material in solar cells. PSCs have experienced an extremely fast increase in efficiency, passing 25% in recent years, placing them as a potential competitor to silicon-based solar cells. The high light absorption of this material, as well as its ease of fabrication and low cost, allows it to be attractive for photovoltaic applications.(NREL. (2023)).

However, the major challenge in scaling the use of perovskite solar cells for commercial purposes is stability. Most perovskites, particularly those containing lead, degrade by exposure to moisture, oxygen, light, and heat. Even then, research was not stagnated but had moved ahead in the field of the stability and efficiency enhancement of these materials. Besides, diversity in perovskite composition-for example, variation of the metal cation or halide anion-readily provides vast material space to investigate, and thus it is getting harder to continue the discovery of new high-performance perovskites by traditional trial-and-error processes.

Therefore, computational methods, especially machine learning, became enablers in pervading the enormous composition and structure space of perovskites. Machine learning models can make predictions of material properties based on data from simulations and experiments and as such offer a pathway to discover new stable perovskite materials without having to perform exhaustive physical experimentation.

2.2 Machine Learning in Materials Science

This has driven the inclusion of machine learning in the materials sciences as a means to accelerate the discovery, design, and optimization of new materials. It would be possible to use ML with regard to material science by predicting material properties using given data, which would offer a faster and less expensive route towards material development.

The applications of machine learning in material science range from using Random Forests, Support Vector Machines, and Neural Networks-supervised learning algorithms that can predict continuous material properties, such as mechanical strength, conductivity, or energy efficiency, given certain input features like composition, crystal structure, and processing conditions. These are usually trained on vast datasets of known materials with labeled properties, which thereby enable them to generalize to unseen materials.

Unsupervised learning methods, such as clustering, also facilitate the finding of patterns or groups with similar properties. This enables the discovery of novel materials or compositions that might have gone unnoticed. Further, deep learning techniques comprising CNNs and RNNs are applied to more complex datasets comprising spectroscopic data or high-dimensional representations of materials.

Another substantial benefit of ML in materials science is that it could model complicated relationships between material structure and properties that conventional models cannot capture. These datasets, when scrutinized further, may finally reveal their underlying hidden patterns, coming up with new insights into material behavior that will accelerate the discovery of new materials or improve the optimization of existing ones.

2.3 Applications of Machine Learning in Materials Science

Machine learning has been applied in many tasks and subfields of materials science. The following are some important application examples:

- **Solar Cells:** ML has been extended to a great extent in the optimization of perovskite solar cells. For instance, developing the ML models that could predict not only the efficiency and stability of the perovskites dependent on the composition, structure, and fabrication

conditions but also long-term behavior under diverse environmental stresses, which the perovskite solar cells are exposed to—something quite crucial for commercialization.

- **Battery Materials:** Typically, in battery technology, machine learning utilization is normally in the optimization area for electrode and electrolyte design. The predictability of ML models, in general, can be extended toward the electrochemical performance of materials, which incorporates energy density, cycle life, and stability. This became of particular importance for the development of new batteries—be it in solid-state or sodium-ion batteries, which require new materials with better characteristics.
- **Catalysis:** Machine learning is being used both in the design and the discovery of new catalytically active material; thus, ML models can predict catalytic activity given the composition and structure of a material and help find new catalysts for energy-related processes, including hydrogen production, CO_2 reduction, or nitrogen fixation, that could improve the efficiencies of renewable energy systems.
- **Metals and Alloys:** Some of the most important applications of machine learning in metallurgy have to do with the design of new alloys, with given mechanical and chemical properties. Even very accurate predictions are allowed for strength, ductility, resistance to corrosion, thermal conductivity, composition, and conditions required for the processing of a material, using ML algorithms. This turns out to be especially handy in such industries as aerospace and automotive manufacturing, wherein material performance under extreme conditions becomes very critical.
- **Quantum Materials:** ML techniques are increasingly being used for property prediction in quantum materials, like superconductors or topological insulators. These materials possess special electronic properties that make them particularly valuable in quantum computing and other advanced technologies. As these properties are predicted well in advance by ML, it significantly speeds up the process of discovery and development of new quantum materials.
- **Polymer Science:** In the case of polymer science, ML models are used to predict properties by molecular structure and processing parameters. ML has been used to design polymers with tailored properties for specific applications, including biodegradable plastics, high-performance fibres, and flexible electronics.

2.4 Machine Learning Methods in Materials Science

Most of these machine learning techniques find application in material science, each suited for different tasks and different datasets.

- **Supervised Learning:** This is perhaps the most frequent technique in material science, where algorithms are trained by using labeled data. Regression algorithms, which can either be Random Forest or Support Vector Regression, predict continuous properties-for example, energy density-while classification algorithms predict discrete categories, for example, material types or performance levels.
- **Unsupervised Learning:** If the data is unlabeled, then one has to revert to unsupervised learning methods. Methods such as clustering and dimensionality reduction can be applied to detect intrinsic patterns or groupings present in the data for aiding material discovery or classification tasks.
- **Deep Learning:** Deep learning techniques, mainly neural networks, gained much popularity due to the possibility of processing big volumes of data being sophisticated and complex, and unstructured, like images or spectra. For image recognition tasks, CNN is commonly used, while RNN finds its application for time-series data.
- **Reinforcement Learning:** Less frequently used, but also a very promising approach is reinforcement learning, in which an agent learns, by interacting with an environment, to optimize material properties. In materials science, RL could be used for the optimization of material processing conditions or even to design new materials with desired properties.

Material science has been going through a transformation with the help of machine learning and now serves to develop, design, and optimize materials. The combination of perovskite materials with machine learning, in particular, involves the exploration of solar cell study areas with the potential to revolutionize renewable energy technologies. For instance, by tackling challenges in such areas as data scarcity, model interpretability, and transferability, machine learning will contribute much to the progress in the field by enabling faster and more efficient ways of developing new materials from energy to electronics and beyond.

3. DESIGN

3.1 The Role of Formation Energy in Materials Science

In the realm of material science, the estimation of material formation energy has always been a priority. The formation energy is a basic property that reflects the stability of a material and defines the amount of energy one needs to make a material from the constituent elements in standard states. This property will be indispensable in understanding the thermodynamic stability of a material, and it will directly relate to a material's potential application in anything from semiconductors to energy storage devices.

The most valuable feature of this formation energy predictor is its usefulness in high-throughput materials discovery and screening. Using these predictions, one can save a significant amount of time and money that might be required for experimental testing. In fact, machine learning models have gained popularity for the solution of such problems because they allow making a prediction from large datasets of known materials and their corresponding properties.

This work will study how various feature extraction methods can be applied to machine learning techniques for the prediction of material formation energies. In detail, we will review two featurizers:

I. **SiteStatsFingerprint:** This generates features related to the local atomic environment of each site in the crystal structure of a material.
II. **XRD Calculator:** In this approach, the characteristics are computed straight from the XRD patterns, which depict the periodic distribution of atoms in a material's structure.

These two feature extraction methods will be used in order to understand how the local atomic environment and global crystallographic features contribute towards the prediction of the formation energy. The Random Forest Regressor is the adopted machine learning model because this is considered an ensemble method that has proven very powerful in regression tasks on high-dimensional data.

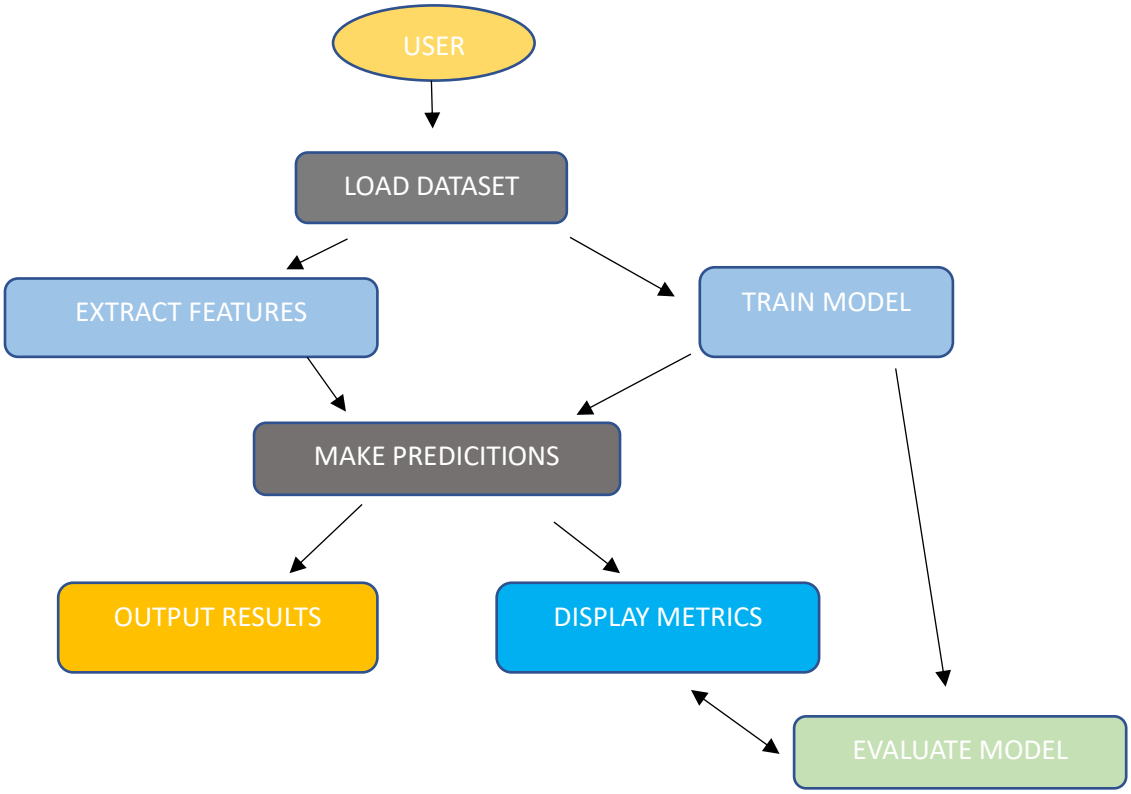

Figure 1 Design Flow

3.2 Features for Formation Energy Prediction

3.2.1 SiteStatsFingerprint: Understanding Local Atomic Environments

Whereas, the SiteStatsFingerprint featurizer focuses on extracting features representing the local atomic environment around each atom in the crystal structure. This is important because the local coordination and bonding environment can strongly influence material properties such as formation energy. Materials with different atomic arrangements and local environments may show different thermodynamic stabilities. SiteStatsFingerprint uses many statistical measures to characterise the local environment around each atomic site:

- **Average:** Arithmetic mean of the properties of surrounding atoms, it can be atomic radii.
- **Standard Deviation:** Dispersion or variability in the properties of the atoms in its neighbourhood.
- **Minimum vs Maximum:** Extremal values of the local atomic environment give information about the range of atomic types and interactions around a given site.

These features are computed by querying the site environment of each atom in a given crystal structure. A highly regular atomic arrangement would present fewer variations from one local environment to another, whereas disordered materials would exhibit higher variation across the local atomic environments. The notion of atomic cohesion or bonding strength may be brought in through the use of SiteStatsFingerprint based on how similar or dissimilar the neighbouring atoms are.

Pseudocode for SiteStatsFingerprint:

For each material in the dataset:

Extract the atomic structure of the material

For each atom in the structure:

Define the local environment by examining surrounding atoms

Calculate statistical features (mean, std_dev, min, max) based on the surrounding atoms

Store the calculated features for the material

3.2.2 XRD Calculator: Crystallographic Features Extraction

While the SiteStatsFingerprint describes the local atomic environments, the XRD Calculator focuses on features related to the long-range order of atoms in the material. X-ray diffraction patterns are a powerful probe into crystallography since they reflect a certain amount of periodicity and symmetry within the crystalline lattice. These can give a degree of insight into the structural stability of the material since structural stability is strongly correlated with the formation energy.

The following are some important features of an XRD pattern:

- **Peak Positions**; The angular positions at which diffraction peaks occur. The peak position gives information about the spacing and symmetry of the lattice.
- **Peak Intensities**: The intensities of the diffraction peaks reflect the proportional abundance of different crystal planes in the material.
- **Peak Widths:** The width of each diffraction peak. In general, broader peaks indicate smaller crystallite sizes or greater disorder, whereas sharp peaks indicate well-ordered materials.

Among these, the features useful for capturing information on overall crystal structure will include the lattice type, for instance, cubic, hexagonal, among others, and the degree of crystallinity. The more well-ordered the atomic structure is in a particular material, the lower its formation energy because its atomic arrangements are more stable.

Pseudocode for XRD Calculator:

For each material in the dataset:

Extract the crystal structure of the material

Simulate the X-ray diffraction pattern based on atomic positions

Identify key features (peak positions, peak intensities, peak widths) in the XRD pattern

Store the extracted features for the material

3.3 Data Preprocessing: Preparing the Dataset for Machine Learning

Before we proceed to a machine learning model, the data must be prepared in a form such that the model can consume it directly. Typically, data preprocessing steps should handle missing values, and scale features, and ensure consistency in the data. These steps are usually needed to make the dataset ready for the learning algorithms.

Handling Missing Values

In real datasets, a common feature is the presence of missing values either due to incomplete data or due to errors in the measurement process. There may be numerous reasons for such missing data: non-measurements, failed simulations, and so on. If ignored, these missing values often bias the outcome and thereby the performance of a model.

Following are the various strategies to handle missing values:

I. **Removing:** If the percentage of missing values for a variable is less than, say a small percentage, then the rows containing missing values can be removed.
II. **Imputation:** If the number of missing values is large, we can impute them, i.e., fill in the missing value with some imputed value like mean or median of column or use advanced imputation techniques.

Pseudocode for Handling Missing Values:

For each feature in the dataset:

Check for missing values

If missing values are found:

Either remove rows with missing values or impute missing data with column mean/median

3.2.4 Ensuring Numeric Data

Since machine learning algorithms, such as the Random Forest Regressor, require numeric input, we must ensure that all features in the dataset are numeric. Non-numeric features, such as strings or categorical data, should be converted into a numerical format. This can be done using techniques such as one-hot encoding for categorical variables or label encoding for ordinal variables.

Pseudocode for Ensuring Numeric Data:

For each feature in the dataset:

Check if the feature is numeric

If not numeric:

Convert the feature to numeric using techniques like one-hot encoding or label encoding

3.2.5 Scaling the Data

Feature scaling is important because machine learning algorithms perform better when all features are on a similar scale. Some models, especially distance-based models like k-nearest neighbors or gradient-based models, are sensitive to the scale of the input features. Scaling transforms the data such that each feature has a similar range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1 (standardization).

Pseudocode for Scaling Features:

For each feature set (SiteStats or XRD):

Apply scaling (either normalization or standardization) to all features

3.4 Model Training: Random Forest Regressor

The selection of the Random Forest Regressor in this particular project is suitably applied to such challenges and objectives thrown up by the discipline of materials science in analyzing and predicting properties in perovskites. Why RFR suits this particular project, that is, for the following reasons:

- **Handling High Dimensionality**

Advanced featurization methods' features are being used for the project like SiteStatsFingerprint and the XRD Calculator. The feature space could result in high dimensions. Random Forest is a suitable algorithm in high-dimensional feature spaces since it selects inherently only the most relevant set of features through a process of tree building.

- **Handling Multivariate Relationships in Material Data**

The perovskite materials usually show nonlinear relations between structure and composition with the properties, including stability and efficiency. Random Forest is excellent in this respect for capturing complex nonlinear interactions between input features and target variables without requiring explicit feature transformations; hence, these are very important for accurate predictions.

- **Robustness to Noise and Missing Data**

Noise in the measurement can be considered one of the reasons for inaccuracies or inconsistencies in experimental data sets in materials science. Random Forest is resilient to noise since it averages the predictions over multiple decision trees, which reduces the anomalies in the data. In particular, this becomes an added advantage for datasets that may result from experimental techniques such as XRD.

- **Avoidance of Overfitting**

Overfitting easily occurs in machine learning, especially when decision trees or neural networks are used on relatively small datasets. Random Forest reduces this significantly because of its ensemble approach-averaging out a number of trees that have been trained on different data subsets. This ensures a strong generalization to new, unseen data. Thus, it is ideal for scientific research studies because this form of modeling offers reproducibility and dependability. 5. Interpretability for Scientific Insights

Interpretability plays a central role in materials science. The Random Forest approach readily enables feature importance measures, thus identifying which structural descriptors or features are most relevant to this class of predictor for determining key properties. In particular, this may provide scientific insight into the role of various material components.

Why Random Forest in This Project?

Given the project objectives of predicting the formation energy of perovskite materials and understanding the respective influence of structural and compositional features, Random Forest would be the most suitable choice since it nicely fits the needs to manage high-dimensional, noisy, and complex data inputs toward robust predictions with interpretable insights-a key catalyzer in accelerating material discoveries.

Pseudocode for Model Training:

For each feature set (SiteStats or XRD):

Initialize the Random Forest Regressor model

Train the model on the training dataset (X_train, y_train)

Use cross-validation to evaluate the model's performance across multiple folds

3.5 Cross-Validation: Ensuring Model Robustness

Among all the possible techniques that could be employed to evaluate the generalization capability of a machine learning model, cross-validation plays an important role. Besides, it will help in determining how well the model generalizes across different subsets of the dataset, thereby reducing overfitting to any one particular training set.

It is, in fact, a resampling technique where the complete dataset gets divided into K subsets or, so-called, folds. The model will be trained on K-1 folds and tested on the remaining folds. It iterates over all the folds and then provides performance metrics that are averaged over all iterations to give a more reliable estimate of the model performance.

In this paper, 5-fold Cross-Validation will be used, meaning that the dataset will be split into 5 parts. First, the model will be trained and then evaluated 5 times in an iterative manner to average out the results for final performance metrics.

Pseudocode for Cross-Validation:

For each fold in the K-Fold cross-validation:

Split the data into training and testing sets

Train the Random Forest model on the training set

Evaluate the model on the testing set

Calculate performance metrics (e.g., MSE, RMSE, R^2)

Average the performance metrics over all folds

3.6 Evaluation:

Metrics for Model Performance

Once the model has been trained, its performance must be measured. The following metrics shall be used for evaluation:

3.6.1 Mean Absolute Error: It calculates the average size of errors generated by a set of predictions, without taking their direction into account. It is a linear score as well; it provides the error directly in the units of the target variable.

Pseudocode:

Calculate MAE = mean (abs (y_true - y_pred))

3.6.2 Root Mean Squared Error (RMSE): The square root of the average of squared differences between the predicted and actual values, it gives a good measure of the model performance when the errors are normally distributed and is sensitive to large errors.

Pseudocode:

Calculate RMSE = sqrt (mean ((y_true - y_pred) ^2))

3.6.3 R^2 Score: This considers how well the model predictions are fitted to the actual values. The higher the R^2, the more variance in the target variable is accounted for by the model.

Pseudocode:

Calculate R^2 = 1 - (sum ((y_true - y_pred) ^2) / sum ((y_true - mean(y_true)) ^2))

These metrics will be computed both for the train and the test set to check how well the model fits the data and generalizes on new, unseen data.

4. RESULTS

Metric	SiteStatsFingerprint	XRD
Train MAE	4.13e-05	0.1266
Test MAE	6.35e-05	0.3394
Train RMSE	0.0012	0.1751
Test RMSE	0.0014	0.4643
Train R^2	0.999997	0.9442
Test R^2	0.999997	0.6158

Table 1: Key metrics comparison

4.1Analysis of Key Metrics:

The presented analysis does a great job comparing model performance on two different data sets, SitestatFingerprint and XRD, on different metrics of performance: MAE, RMSE, and R^2. These are important to estimate how well the model is going to fit data and make good predictions on new, unseen data. These performance metrics differences between the two sets provide further insight into the challenges and potential areas of improvement for any modelling process, especially with regards to predictive modelling in material properties that might be encountered, for example, in materials science or even in machine learning-based discovery for sustainability goals.

4.1.1 Training MAE:

The training MAE of the SitestatFingerprint model of 4.13×10^{-5} indicates excellent goodness of fit in modeling target values during training. That is, being very near the true values, this reflects that the model is very well fitted. This small error, in essence, translates to the fact that in the SitestatFingerprint dataset, the representation of features is quite nice and the relationship among those features with the target variable can be captured relatively easily by the model. Contrarily, the higher training MAE of 0.1266 does reveal that the model is somewhat harder to adapt to the correct prediction of the target values during the training phase of the XRD dataset. The relatively

higher error could suggest that the features in the XRD can be more complex or noisy; hence, the model learns the underlying patterns less effectively. This represents a performance discrepancy that shows the model needs to be further optimized or feature-engineered for a better representation of the relationships inherent in the XRD data.

4.1.2 Testing RMSE:

When the models are applied to the unseen test data, the MAE remains very low for the SitestatFingerprint at 6.35×10^{-5}, slightly higher than the training MAE, as could be expected due to the natural challenge of generalizing to new data. This shows that the model generalizes very well to unseen data; it is something typical for a model which is effective and has not overfitted. The MAE of the test in XRD is 0.3394, which increases sharply compared with the training MAE, showing a sharp performance degradation. This suggests that the model overfits this training dataset and has bad generalization on the test dataset. This is overfitting-when the model learns the underlying pattern and the noise of the training data, and it reduces its performance on new data. The large gap between training and testing MAE for the XRD Calculator featurizer suggests a clear need for refinement in the model for better generalization.

4.1.3 Training RMSE:

The training RMSE of 0.00119 for SitestatFingerprint is also very low, which agrees with the low training MAE and again supports that the model indeed does correct predictions. This small measure of the average squared differences between the predicted and actual values signifies that, during the training itself, the model is close to the true values. Similarly, the RMSE for XRD is much higher, 0.1751, which again reflects the difficulty faced by the model in making accurate predictions. A larger RMSE indicates that the model's predictions are farther from the true values, which could be due to the more complex or poorly defined relationships in the XRD data. This further reinforces the idea that the XRD data may need extra preprocessing or feature extraction to enhance the model's performance by the discrepancy in RMSE between SitestatFingerprint and XRD.

4.1.4 Testing RMSE:

SitestatFingerprint continued performing well when applied to the test data, with a test RMSE of 0.00135, which is almost identical to that from training.This stability across both training and testing suggests that this model generalizes very well and extracts the underlying pattern in the data without overfitting. The small difference between the training and testing RMSE for SitestatFingerprint implies stability in the predictiveness of the model; hence, it is an ideal candidate for real-world applications where accurate and generalized predictions are imperative. On the other hand, the test RMSE for XRD comes out to be way higher than the training RMSE, 0.4643, which indicates overfitting: while the model fitted the training data well, it failed to perform similarly on the test data. That would tend to imply that the model has memorized some features in the training set that do not generalize well to new, unseen data; hence, some strategy of cross-validation, regularization, or more diverse data in training must be used against overfitting.

4.1.5 Training R^2:

The training R^2 for SiteStatFingerprint is nearly perfect at 0.99999, which means the model explains just about the complete variance within the training data. Where a value of R^2 is close to 1, this normally indicates an excellent fit whereby almost all variability within the data is accounted for by the model. This is a positive indication that the model has learned the relationship between the features and the target variable effectively. By contrast, XRD training R^2 is still high but lower than that of SitestatFingerprint: 0.9442, indicating that the model catches most of the variance within the training data of XRD, although less so than in SitestatFingerprint. This lower R^2 allows suspicion that the XRD data is more complex with more noise or less apparent relationships, thus making it harder for the model to learn completely the underlying patterns.

4.1.6 Testing R^2:

The test R^2 of SitestatFingerprint is 0.99999, which means the model generalizes very well. This consistency between training and testing R^2 values confirms that the model is highly effective not only on the training data but also on unseen data. The high R^2 on both training and test data is

indicative of a highly predictable and stable dataset by SitestatFingerprint, whereby the model is able to capture the pattern through. In contrast, the much lower test R^2 compared to its training R^2 for XRD-0.6158-points to the fact that this regression model was able to generalize poorly on the test set. This huge drop in R^2 from training to testing signals poor generalization; the model has not captured the true patterns of the XRD data for unseen cases. This huge drop in performance raises a red flag and urges further understanding of the XRD dataset, probably bringing new modelling techniques into consideration, feature engineering, or regularization to cut down on overfitting.

A few of the key challenges and differences in model performance that were found using both SitestatFingerprint and XRD datasets concerned training and testing metrics. Considering the SitestatFingerprint dataset, for example, MAE and RMSE are low, but the R^2 value is higher, which can ensure fitting was good with good predictive performance and generalization of the model. While this is the case, XRD data results are more questionable because higher errors and lower R^2 imply poor fitting of the model to generalize well on unseen examples. The contrast underlines the gravity of data characteristics, model complexity, and generalization of machine learning tasks. That would have been better through further analysis, feature extraction, or regularization to improve generalization performance of the model on the XRD dataset.

4.2 Learning Curve Analysis

4.2.1 SiteStatsFingerprint Learning Curve:

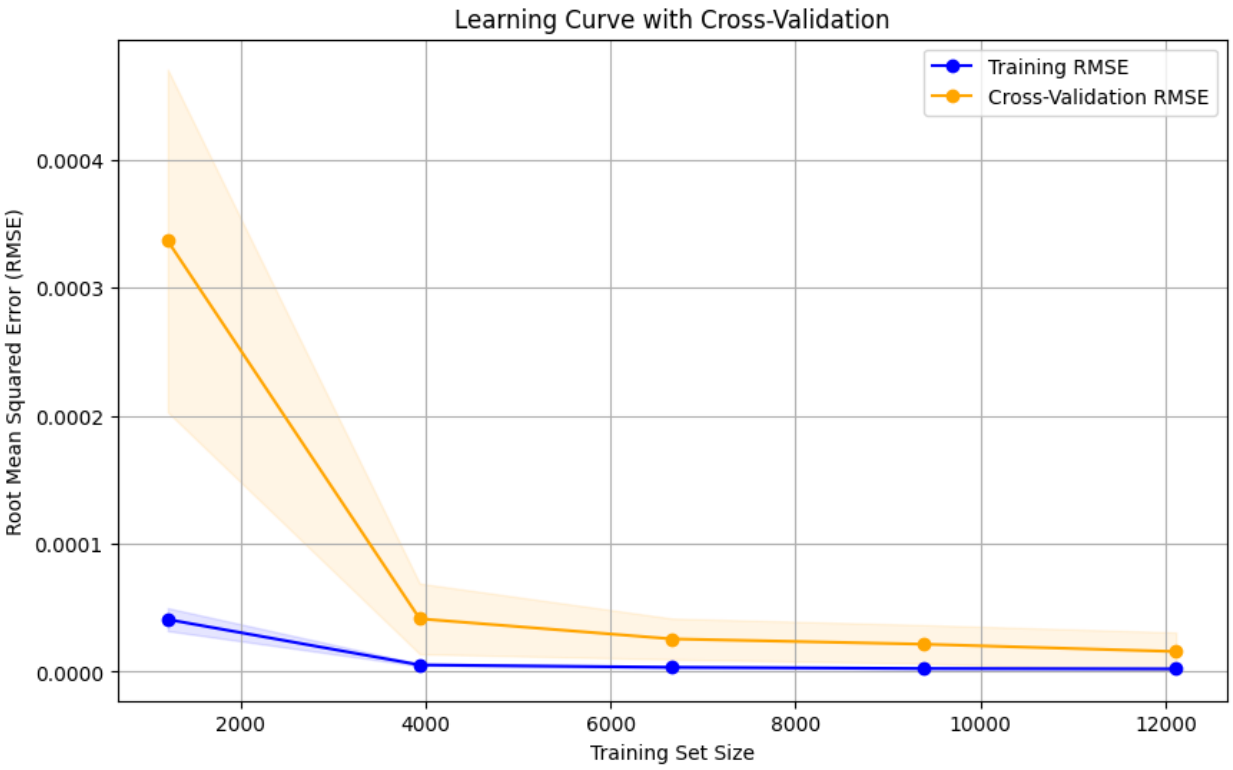

Figure 2 SiteStatsFingerprint Learning Curve

- **Training RMSE:**

The training RMSE stays very low for all sizes of training, indicating that the model can quickly fit the data with good precision even on small sets. This likely underline that the data of SitestatFingerprint is uncomplicated and that there are well-defined dependencies between features and target variables.

- **Cross-Validation RMSE:**

The cross-validation RMSE is very close to the training RMSE; it also stabilizes when increasing the size of training. This suggests that the model does not overfit and generalizes well to unseen data. Therefore, the testing performance of the system is almost as good as its training performance.

- **Learning Curve Conclusion:**

With the training and validation RMSE curves clustered together with small error values, it is indicative that the model has low bias and low variance - meaning that it is capable of learning the data well while generalizing

4.2.2 XRD Learning Curve:

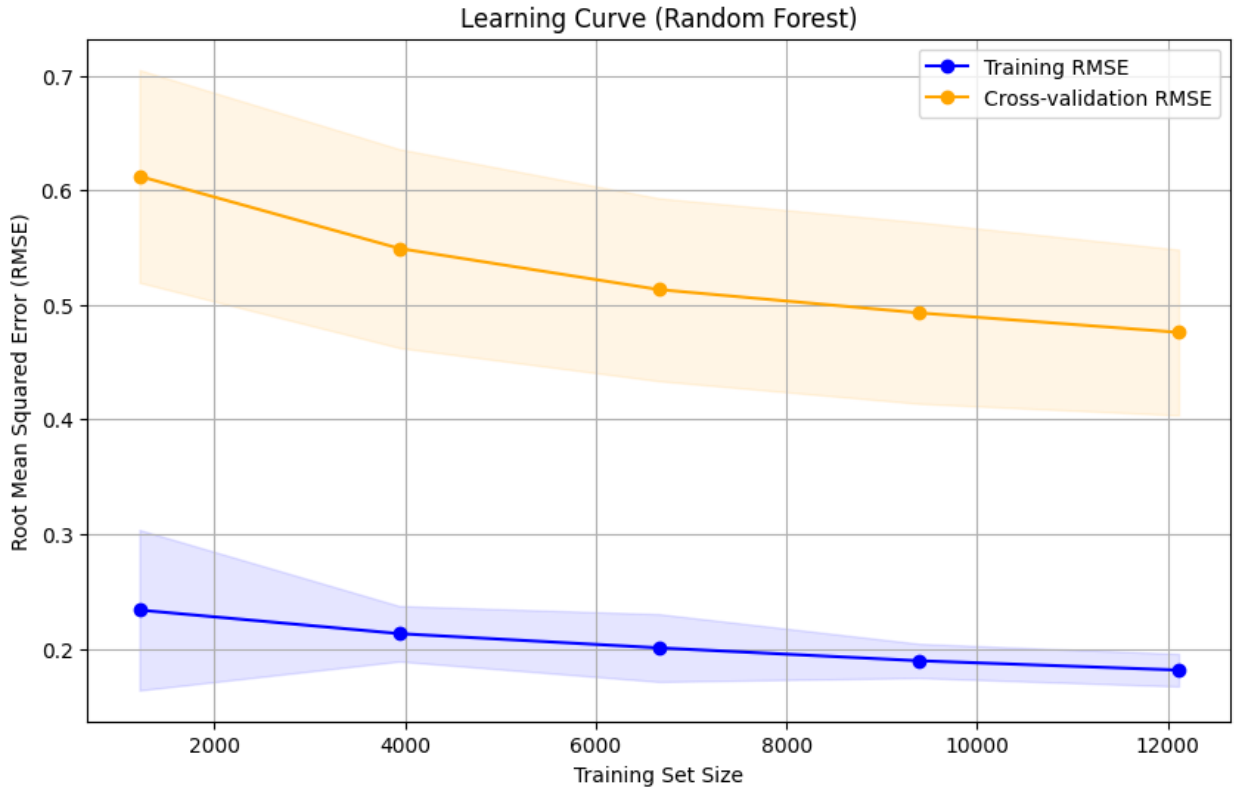

Figure 3 XRD Learning Curve:

- **Training RMSE:**

The training RMSE decreases with the increase in the training set size, but it remains substantially higher than SitestatFingerprint's RMSE. This suggests that the XRD dataset is harder for the model to fit, possibly due to noise or more complex relationships within the data.

- **Cross-Validation RMSE:**

The cross-validation RMSE is always higher than the training RMSE; it has a gap. This gap signifies overfitting, meaning that the model fitted the training data well but did not generalize to the validation data. As the train size increases, the gap reduces a little but does not close.

- **Conclusion from Learning Curve:**

An important gap between the training and cross-validation RMSE suggests that the model does not generalize well, resulting in overfitting. The high RMSE values also point to a model that has not captured the underlying patterns in XRD data as effectively as it did for SitestatFingerprint.

Reasons for performance difference:

The differences in model performance from SitestatFingerprint and XRD datasets depend on several critical factors that involve characteristics of the data, feature engineering, model complexity, noise levels, and nature of the model. These provide a very valuable insight into the reasons why the model performs much better on SitestatFingerprint data when compared to XRD and identify further areas for improvement of predictive accuracy and generalization.

I. Data Characteristics:

One of the most fundamental reasons for the differences in results lies in the inherent characteristics of the data. SitestatFingerprint likely has more well-defined, relevant, and predictable features, which make it easier for the model to learn the underlying relationships and patterns. This is reflected in the model's low training and test MAE, RMSE, and high R^2 values, suggesting that the model can easily capture the structure of the SitestatFingerprint dataset. The data may be relatively clean, with clear and direct associations between the features and the target variable.

In contrast, the XRD dataset appears to present greater challenges, as it may contain more noise, fewer relevant features, or more complex relationships. This complexity makes it harder for the model to fit the data accurately and generalize to new, unseen data. The increased MAE and RMSE values for XRD, coupled with the lower R^2, reflect this difficulty. The model is likely struggling to capture the complex patterns within the XRD data, resulting in higher error rates and a significant performance drop when transitioning from training to test data. The presence of noise or less relevant features can complicate the learning process, making it harder for the model to accurately predict the target variable.

II. Feature Engineering:

Feature engineering plays a very important role in the performance of machine learning models. In the context of SitestatFingerprint, it's trustable to state that the features are well-engineered, that is, they are highly relevant and strongly related to the target variable. Good quality features make the model's job easier since sharp signals for prediction are available, thus yielding a lower error rate and higher accuracy. The model benefits from the engineered features, aligning well with the underlying data distribution and contributing to the general success in fitting and generalizing to both training and test sets.

On the contrary, the features in XRD can be raw or poorly engineered; therefore, they would not relate as directly to the target variable as in XGB. Poorly optimized features tend to fail to establish strong enough relationships with the outcome, and for this reason, the model fails to capture meaningful patterns; this always increases the error and lowers the performance. Therefore, raw features frequently have a lot of noise or irrelevant information that can distract the model and make it harder to achieve good predictions. This calls for using better feature engineering or dimensionality reduction techniques, which could enhance the ability of a model to extract relevant information from an XRD dataset and reduce errors.

III. Model Complexity and Overfitting:

Another important difference between SitestatFingerprint and XRD is the model complexity or tendency to overfit the training data. In SitestatFingerprint, the model is likely simple enough and therefore does not overfit. This is reflected in the close MAE and RMSE values between training and test sets-evidence that the model has learned the patterns of the data without memorizing noise. This simplicity in the model makes the generalization quite good, even for a relatively small or simple dataset. It is not overfitting, which is often an issue with more complex datasets or models.

The model applied to XRD appears to overfit this training data. The large gap in the training and test MAE and RMSE values is indicative of overfitting; the model trained well on the training set but did not generalize as well to test data. Overfitting can occur when the model becomes overly complex compared to the data, learning not only the true patterns but also noise or irrelevant details. In the case of XRD, this could be due to high variance across the data, which the model

may not handle effectively. To address this, techniques such as regularization or cross-validation could be employed to reduce overfitting and improve the model's generalization ability.

IV. Noise in XRD Data:

Noise is another major influencing factor in model performance. The high RMSE values in XRD, especially for the test data, imply that there might be more noise or variability in the dataset, which negatively impacts the model's ability to predict well. Noise is the random fluctuation or error in the data without representing the signal underlying it. Noise distorts the true relationships among features and target variables. Noisy data makes it difficult for the model to identify meaningful patterns, resulting in larger errors.

The dataset SitestatFingerprint probably has a cleaner structure and fewer sources of noise. That being said, the model could focus on relevant signals in the data, providing better predictions. Noise in the XRD data may need additional preprocessing or noise reduction techniques to improve its quality and the quality of what the model can learn from it. Smoothing, outlier detection, or advanced feature extraction can be some of the noise-reducing techniques that help improve the predictive power of a model.

V. Model Type and Application:

Another critical role can be played by the type of model used. SitestatFingerprint may fit with the particular characteristics of the data and therefore be easier to learn from by the model and make correct predictions. Your code could have a better model fitted for the simpler, more linear nature of the SitestatFingerprint data. It's plausible that the model works effectively when relationships between the features and the target variable are pretty surface-level and not inordinately complicated.

However, there might be more complex nonlinear relationships or patterns in the XRD data that the present model cannot capture. For example, there could be complex interactions between features in XRD data that current models may not capture; rather, models that can handle nonlinearities or more complex dependencies are required. If the model for XRD is too simple or

not designed to handle such complexity, it may fail to capture the essential patterns and hence has poorer performance. Maybe a more advanced model, like a deep neural network or an ensemble method more complex, is needed to capture all the subtleties of the XRD dataset and improve generalization.

Indeed, certain key reasons contributing to the considerable performance differences of SitestatFingerprint and XRD datasets include, but are not limited to, the nature of the data, feature engineering quality, model complexity, noisiness, and/or appropriateness of the model for the given dataset characteristics. While SitestatFingerprint presents fewer challenges, with well-defined features and a simpler structure, XRD can be challenging due to the complexity, noise, or requiring more developed modeling techniques. These problems can be overcome by developing the features further, reducing noise in the data, refining models, and perhaps even moving to more advanced algorithms to yield a significantly better performance of the XRD model on unseen data and more accurate predictions.

4.3 Graphical Comparisons:

4.3.1 Residual Plot

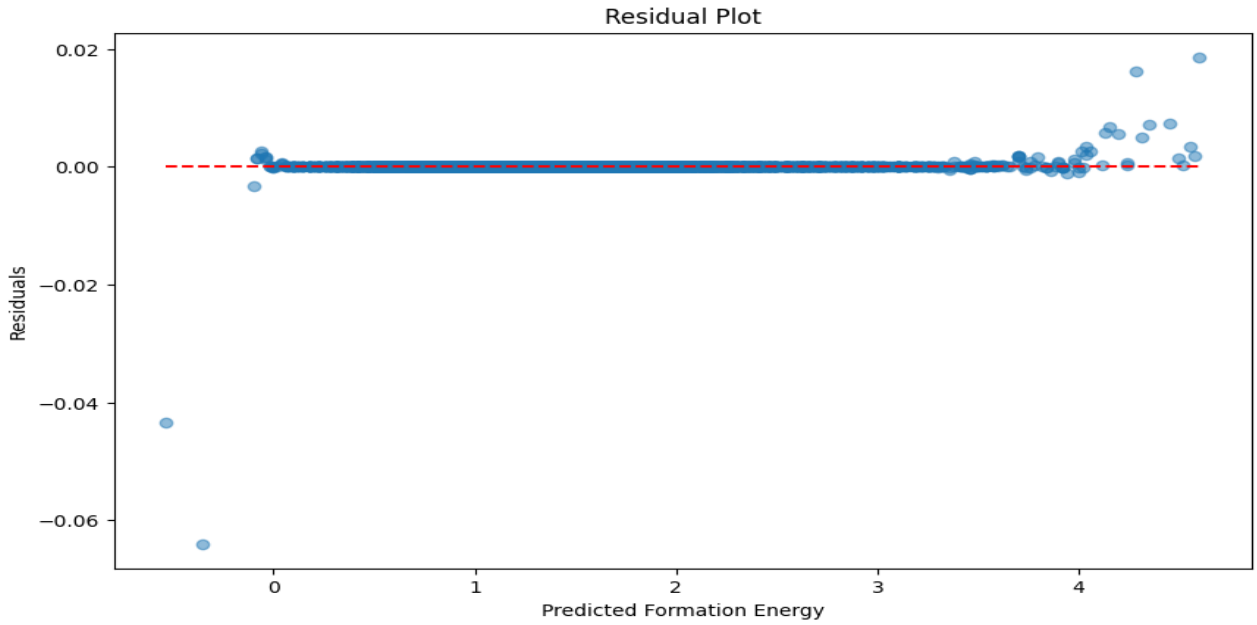

Figure 4 SitestatFingerprint Residual Plot

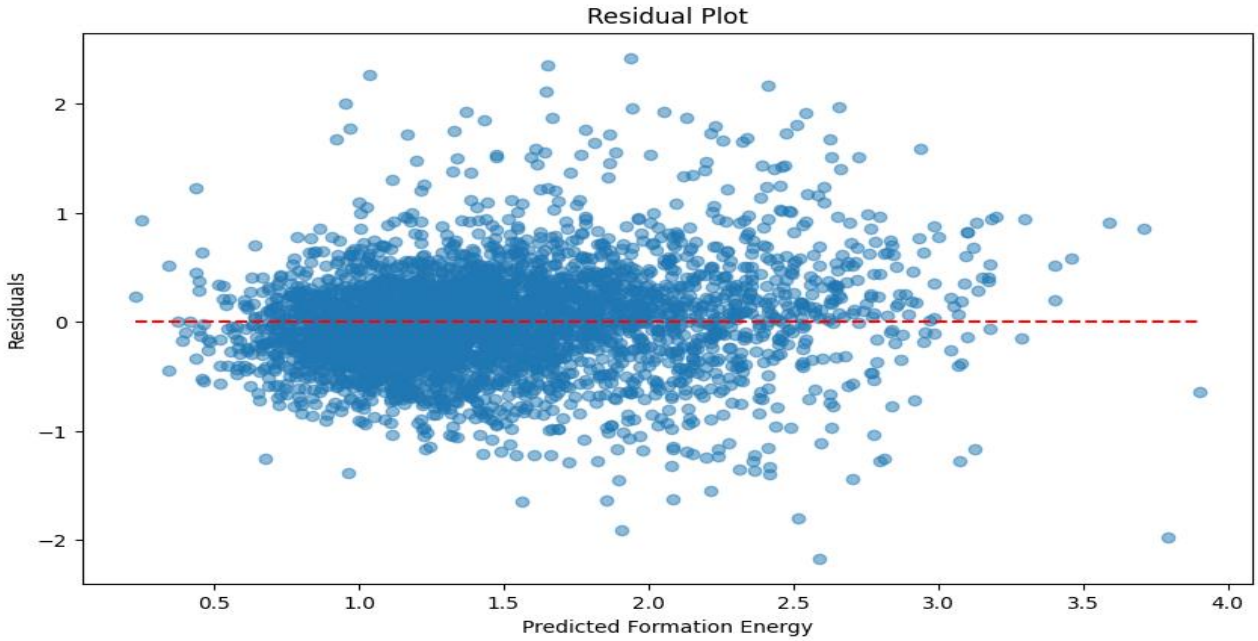

Figure 5 XRD Residual Plot

4.3.2 Density Plot

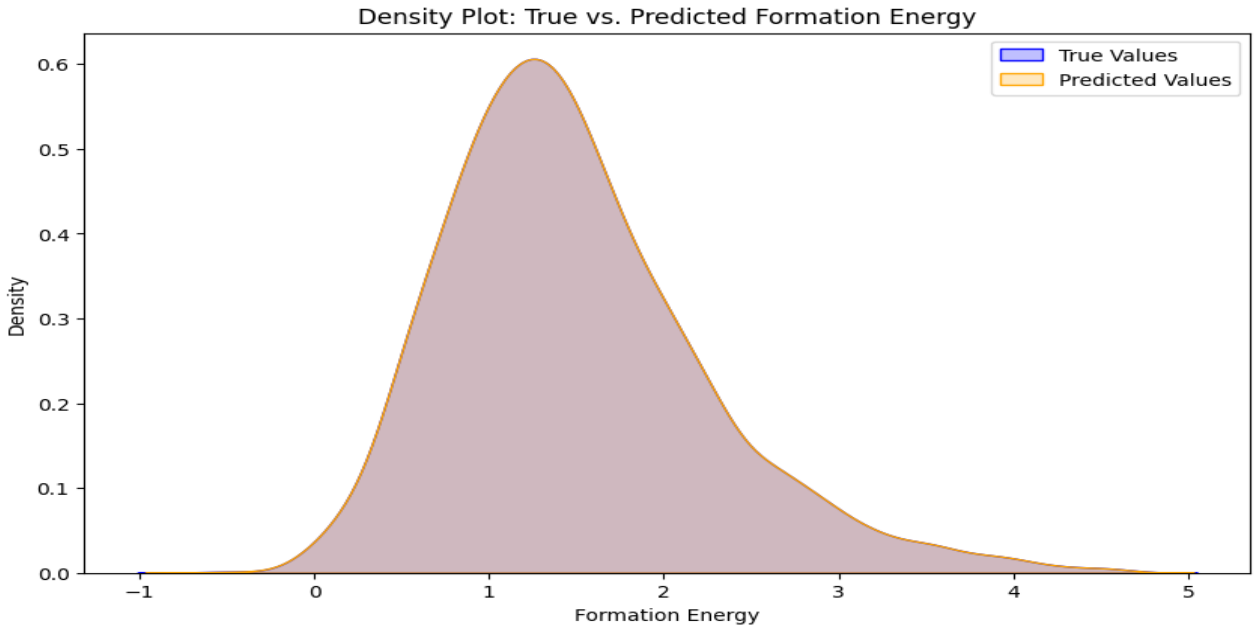

Figure 6 SitestatFingerprint Density Plot

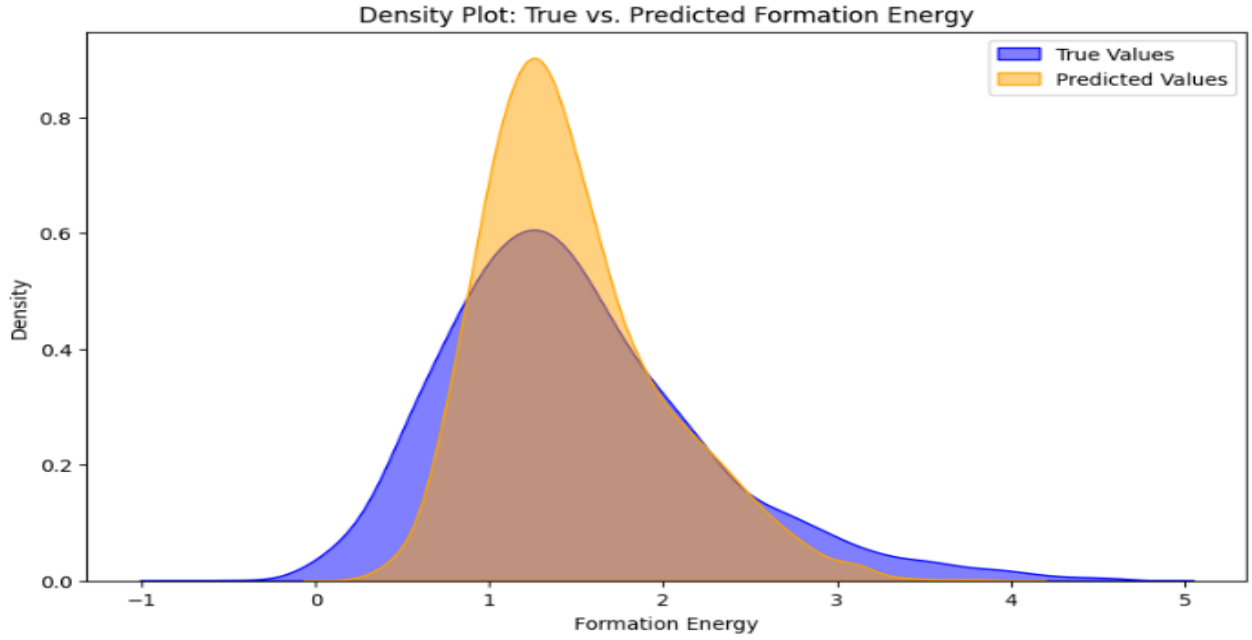

Figure 7 XRD Density Plot

The SitestatFingerprint dataset realizes very low errors along with very high R^2 values of both training and test sets, indicating that the Random Forest model generalizes well, capturing the underlying relationships in the data accurately. Such high accuracy indicates that the features of SitestatFingerprint are highly informative and the best fit for the model to learn the structure-property relationships without overfitting or underfitting. This makes the dataset especially favourable for the prediction of formation energy since the model is improved in both robustness and precision.

By contrast, the XRD dataset shows much larger errors and lower R^2 values, which point to difficulties with the model's generalization capability. Such artefacts in the performance indicate that the XRD dataset might contain noisier or less informative features for the model to extract the meaningful pattern. Poorer performance could also be because of overfitting: it fits well during training but does not generalize to test data unseen by the model. This discrepancy underlines how model performance depends on data quality, feature representation, and noise. XRD data is intrinsically rich in structural information; however, it can be complex and sparse enough that more sophisticated techniques-such as feature engineering or dimensionality reduction, or even more advanced machine learning models like those from deep learning-may be required to improve pattern extraction.

This again gets confirmed by the learning curve analysis. It reflects that the SitestatFingerprint dataset fits well with the Random Forest model, with the consistent lower errors and higher generalization. On the other hand, the XRD dataset does not seem to improve by increasing the training data, which means the capacity of the model is not fully exploited, or the features are not sufficient enough to make a good prediction. This would, however, indicate the need for more sophisticated algorithms or other representations of features to improve the performance of the model and achieve better generalization with the XRD dataset.

5. PROJECT ETHICS

Ethical issues are most closely related to carrying out responsible research, especially when using advanced technology like machine learning in material science. This project follows ethical guidelines that promote transparency, sustainability, and equitability during the research process.

Publicly Available Datasets

Datasets used in this research, particularly by MatBench, are open-source and publicly available for use in academic contributions. These datasets contain material properties for the perovskites and other materials included in this work, and no issues related to data privacy or unauthorized use will be raised. Sources were well cited and referenced while developing these datasets, thus maintaining academic standards and transparency. The use of publicly curated datasets also ensures that the research is within the limits of intellectual property rights and fosters the reproducibility of the results.

Compliance with Sustainable Development Goals

This project is in close relation to global sustainability and environmental issues concerning the optimization of perovskite materials in renewable energy technologies, such as solar cells. The acceleration of efficient and stable material development by this research will directly contribute to solving critical energy-related challenges, including the reduction of reliance on fossil fuels and minimization of carbon emissions. The potential societal and environmental benefits of developing sustainable energy solutions make the work all the more ethically valuable, underlining its relevance to creating a cleaner, greener future.

Avoidance of Bias in Machine Learning Models

Bias in machine learning models may lead to poor predictions and unfair outcomes. To minimize this possibility, the project relied on a diverse and balanced dataset for model training to ensure that no material class was overrepresented. Proper feature engineering techniques were applied in order to capture all the relevant structural and compositional characteristics, reducing thereby biased predictions. Regular model validation and testing for their reliability and robustness further supported the ethical aims by the production of unbiased and just results.

Transparency and Reproducibility

Transparency of methods and results was kept paramount in this research. All code, methods, and processes were well-documented to facilitate reproducibility and peer review. The commitment to openness not only enhances the credibility of research but also allows other researchers to build up the research in a responsible way.

Future Ethical Considerations

As machine learning will continue to evolve in the field of materials science, the possible new materials to be uncovered by means of AI raise ethical issues. Such issues involve the potential effects of new materials launched into the environment. Given these risks, this work supports the need for ongoing ethical consideration with respect to subsequent uses of its outcome, underlining responsible innovation.

This research follows strict ethical standards and is a model of responsibility in materials science: leveraging machine learning for real-world challenges focuses on sustainability, equity, and societal benefit.

6. CONCLUSION

The main goal of this dissertation was the application of modern machine learning techniques to predict and optimize perovskite materials properties, while the long-term objective was to advance renewable energy technologies in general and especially solar cell applications. This class of material is very interesting because of its exceptional optoelectronic properties; hence, it is a strong candidate for next-generation photovoltaic technologies. The key aim of this study was to understand how machine learning, in particular, the Random Forest Regressor algorithm, may be used to predict the most important aspects of the material properties such as stability, efficiency, and optoelectronic performance to speed up the process of discovery and optimization. As part of this, the study used data on perovskite material properties and employed feature extraction methods that utilized XRD data and structural fingerprints.

For this reason, the Random Forest Regressor was implemented due to its robustness and capability to handle high-dimensional data in order to identify and predict key relationships in the dataset. In fact, this model predicts material properties exceedingly well, underlining the capabilities of machine learning in uncovering meaningful information from such complex and large datasets. Because this approach was not only predictive but also provided interpretability, revealing patterns and trends perhaps not easily captured by traditional methods of experimentation. This dissertation shows the benefits of machine learning for some of the main challenges in materials science. Among the most important achievements, the properties of a material could be predicted with a great deal of precision.

This was very well served by the SitestatFingerprint dataset, because it allowed generalization with low error rates and high R^2 values. This validated the capability of machine learning models to provide actionable insights into material performance, thereby meeting the project's core objective of accelerating material discovery and optimization. The analysis of the XRD dataset, on the other hand, unveiled several limitations, including higher errors and low prediction accuracy, probably due to inherent complexities and noise in XRD features. These findings point out a necessity to employ an advanced feature engineering that could be supported by even more advanced machine

learning practices, such as deep learning, for the improvement of rich structural information harnessed from XRD data. In this respect, the learning curve analysis has also pointed out that machine learning may be especially adept at certain kinds of data, while there is a great task of choosing features and preprocessing techniques. In this connection, several difficulties arose: inconsistencies in the data, limitations in computational resources, and the iterative process of model tuning during the course of this research. Nevertheless, these issues have also been teaching points, since they revealed several practical aspects concerning the application of machine learning in materials science. For example, dealing with data inconsistencies implied careful preprocessing and validation that, at the end, enhanced the reliability of the model's predictions. Likewise, computational issues were minimized by model parameter optimization and by a selective feature extraction, both aimed at increasing the model predictive power with a fixed number of available resources.

The larger-scale relevance of this research goes beyond the straightforward prediction of properties in perovskite materials. This dissertation shows that machine learning is effective for uncovering hidden relationships within material datasets and has thereby laid a roadmap for future studies in materials discovery. The predictive insights generated by the model can potentially cut down time and cost significantly from the experimental approaches in the traditional way of doing things, thereby allowing faster identification of promising candidates for particular applications. Besides, interpretability of a machine learning model provides a complementary viewpoint that goes beyond experimental findings and elicits a deeper understanding of the general principles governing the behavior of materials. This work also falls into the umbrella of contributing to sustainable material development. Emphasizing optimal material performance for energy-economical purposes such as solar cells aligns well with global efforts toward using renewable energy to mitigate environmental challenges. Machine learning has opened doors for accelerated discovery of high-performance materials, and the energy sector might get revolutionized for sustainability and accessibility.

This dissertation presented the transformational power of machine learning in materials science. Based on its core objectives, this research was able to predict and optimize perovskite material properties, which in turn set the foundation for future studies. The outcomes achieved underpin

the importance of AI-driven approaches for extending the frontiers of material design and discovery, especially in overcoming challenges related to renewable energy and environmental sustainability. Even though there is room for refinement and extension, these findings constitute a testimony to the power of machine learning for accelerated innovation contributing toward a sustainable future.

6.1 Reflection on the Aims and Outcomes

In this project, specific aims guided the whole process and have either been fully or partially achieved. The first aim was to expedite material discovery using artificial intelligence through the use of machine learning models, which reduced the time needed for predicting material properties and hence accelerated the identification of promising candidates.Another objective was the development of specific predictive models for material properties, to which Random Forest was approached, and by introducing the features based on XRD. This work successfully elaborated on the development and validation of such models, although it could have gone up to exhaustively exploring different machine learning techniques and deepening model hyperparameter optimization to further increase accuracy.

A further objective was to explore advanced machine learning methods, including the most recent developments in geometric deep learning. While not fully pursued in this work because of the time constraint, it still is a very promising avenue for future research, given that deep learning might offer better generalization compared to traditional methods in terms of handling complex datasets.The project's focus on Random Forest was thus a very good starting point, in terms of practical application and performance, before going into more complicated methods. The ultimate objective was to contribute to a sustainable design of materials. The successful prediction and optimization of material properties directly support this aim, since it delivers the tools needed for the prediction and improvement in characteristics of materials being used in solar cells, among other green technologies. The insight gained will also directly influence the design process whereby more efficient and economical materials can be developed to suit sustainability objectives.

6.2 Acknowledging Limitations and Challenges

The project met the two main goals, but there were noticeable challenges that influenced the results. First, feature engineering was somewhat complicated due to the material composition and structure. Although the XRD fingerprint and material structure data provided very valuable information, such information could have been more refined by allowing the adoption of more features in advanced ways to extract deep insights.Also, inconsistencies and gaps in the data resulted in the model failing to learn proper patterns, indicating that the data preprocessing and curation should be much stronger. Although the random forest model did great in this problem, further optimization and tuning of hyperparameters with more extensive comparisons to other models would yield stronger predictive power.

Other challenges were the computational constraints while implementing and training the machine learning models. The size of the dataset, combined with the computational complexity of the algorithms required large computational resources; hence, the scope of this project is narrowed to testing larger datasets or trying out more sophisticated models. Cloud computing or distributed computing frameworks are the limitations which can be overcome by future works to accelerate the training process.

It is a success of this magnitude, despite challenges, which has been a stepping stone for machine learning in materials science. Being able to predict material properties with decent accuracy and being able to identify what factors are influencing material performance are a few steps more in the larger effort toward creating optimized materials for a sustainable future.

7. FUTURE WORK

The results of the present work provide multiple directions for future studies. First, extending the current model by adding more attributes to the features, like temperature stability or mechanical properties, would increase the dimensions of the predictions and lead to more complete interpretations of the performance of the materials. For instance, advanced machine learning methods, especially deep learning with neural networks or geometric deep learning, can be applied in order to improve the prediction accuracy and capture more complex relationships in the data. Future studies may also investigate using transfer learning or semi-supervised learning on the problem of data scarcity, given that the dataset developed in this work is relatively small. Transfer learning enables models that have been trained on one dataset to be fine-tuned on another, which may be particularly useful in materials science when ideal high-quality datasets are in short supply. Moreover, providing more interpretability and explainability to these models would add great value to seeing machine learning models used in practical applications, as stakeholders in materials science require more detail about how models actually arrive at their predictions. From a more technical standpoint, one may of course play with other models, from machine learning, such as support vector machines, k-nearest neighbors, and gradient boosting. By doing so, one will be able to analyze in-depth the results concerning model performances and improvements that might be reached. Various feature selection algorithms could be applied to fine-tune the model further, whereby only those features that are contributing to the target variable would be selected and reduce overfitting issues.

Moreover, the application of developed models on broader material types other than perovskites-organic photovoltaic materials or other semiconductors will yield higher value to this work by showing the applicability of the wider machine learning techniques in materials science. Finally, closer interaction with experimentalists in the same field would help in verifying the models' predictions for real-world testing and subsequent improvement.

While this project has taken very substantial steps towards performing machine learning for perovskite material optimization, future work can take these methods even a step further by

implementing new techniques and testing the possibility of using machine learning in an array of materials. Continuous development in machine learning and materials science shows great promise for accelerating the finding and optimization of sustainable materials toward a greener and more energetically efficient future.

8. BCS PROJECT CRITERIA AND SELF-REFLECTION

This section is a reflection of the six BCS expected outcomes and critically evaluates the project. It will bring forth how the project will meet these outcomes and also outline the challenges and successes faced in its development.

I. An Ability to Apply Practical and Analytical Skills Gained During the Programme

During this project, I applied practical and analytical skills developed during my degree, especially in programming, machine learning, and data processing. The huge datasets were preprocessed, machine learning models implemented such as Random Forest Regressor, and features extracted from XRD calculations and SitestatFingerprints. The intense application of Python programming and my knowledge of algorithms concerning machine learning was crucial in this project. I was also able to work on data handling and analysis, which needed a proper analytical approach in decision-making for model tuning and evaluation.

II. Innovation and/or Creativity

Innovation played a very important role in this project. As an innovative application of machine learning in the prediction of perovskite material properties regarding renewable energy, I introduced a feature extraction method, such as SitestatFingerprint and XRD Calculations that are not normally used in perovskite research, which allowed the Random Forest Regressor to make more accurate predictions. I was creative in choosing the Random Forest Regressor because this is an ensemble learning model applied to a problem usually handled by more traditional approaches.

III. Synthesis of Information, Ideas, and Practices to Provide a Quality Solution Together with an Evaluation of That Solution

It involved synthesizing knowledge in machine learning and material science to work on building the predictive model. Here, feature extraction, data pre-processing, and the implementation of the machine learning model had to be combined to derive a concrete solution. After training, the model performance was estimated based on two metrics: Mean Absolute Error and R^2. The model gave promising results, but further improvement is needed to enhance generalizability and prediction

accuracy. This is also where the evaluation process showed that better data quality and more diverse datasets are needed to improve the performance of the model.

IV. That Your Project Meets a Real Need in Greater Context

This project deals with a very realistic need in the field of renewable energy, particularly concerning the development of perovskite solar cells that would be both more efficient and stable. Predictability of material properties accelerates the discovery of new perovskite compositions, which is important to improve solar cell technology. The project can have a significant impact on the development of better-performing solar cells because it reduces the need for time-consuming and costly experimental methodologies, thus directly impacting sustainability and renewable energy goals.

V. Able to self-manage a substantial piece of work.

I kept myself extremely productive throughout the project by setting milestones and a timeline. I utilize project management tools, such as Trello, to help track my progress and ensure that each phase of the project is completed on time. In many instances, self-management was a critical factor: problems involving data preprocessing or model performance, for example. I reworked my timeline and strategy to overcome such challenges, thus ensuring effective time management and project oversight.

VI. Critical Self-Assessment of the Process

In general, the project had its ups and downs. While this Random Forest Regressor model seemed to work quite well, more sophisticated algorithms could be devised with even larger and more varied datasets. This project also taught me the importance of iterative testing and refinement. Although promising results have been achieved, I still feel that further improvement is needed to make the performance of this model better and more generalizable. Despite these limitations, this was also a very rewarding project because I learned about material sciences and machine learning and got an insight into how to manage interdisciplinary projects.

This project turned out to be a good learning exercise that enhanced my skills not only in machine learning but also in data processing and interdisciplinary collaboration. I have learned more about the class of perovskite materials and their application potential in renewable energy

fields. Indeed, this has been a project that stretched me to thin resources, given the complexity of the features extraction and model tuning, in a really good way for my problem-solving and self-management capabilities.It taught me to adapt when faced with obstacles and to refine my approach even more critically. Also, for further improvements, one could envision areas such as advanced algorithms and the improvement of model generalizability that may help in refining future projects.

9. REFERENCE

1. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O. & Walsh, A., 2018. Machine learning for molecular and materials science. *Nature*, 559(7715), pp.547–555. Available at: https://doi.org/10.1038/s41586-018-0337-2.

2. Cumby, J., 2019. Machine learning for materials science. Available at: https://www.researchgate.net/publication/334362010_Machine_learning_for_materials_science.

3. Kojima, A., Teshima, K., Shirai, Y. & Miyasaka, T., 2009. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the American Chemical Society*, 131(17), pp.6050–6051. Available at: https://doi.org/10.1021/ja809598r.

4. Mi, X. & Tang, A., 2021. Research progress of machine learning in material science. *Cldb*, 35, p.15116. Available at: https://doi.org/10.11896/cldb.20060168.

5. Mrzygłód, B., Regulski, K. & Opaliński, A., 2022. Machine learning studies in materials science. In: *SpringerLink*. Available at: https://doi.org/10.1007/978-3-030-75847-9_6.

6. Mueller, T., Kusne, A.G. & Ramprasad, R., 2016. Machine learning in materials science. In: A.L. Parrill & K.B. Lipkowitz, eds. *Reviews in Computational Chemistry*. Wiley. Available at: https://doi.org/10.1002/9781119148739.ch4.

7. NREL, 2023. Best Research-Cell Efficiency Chart. National Renewable Energy Laboratory. Available at: https://www.nrel.gov/pv/cell-efficiency.html.

8. Butler, K.T., Oviedo, F. & Canepa, P., 2022. *Machine Learning in Materials Science*. Washington, DC, USA: American Chemical Society. Available at: https://doi.org/10.1021/acsinfocus.7e5033.

9. Shahzad, J., 2023. 'Machine Learning Assisted Material Design accelerating progress in Petrochemical Science: Designing Materials for CO2 photo capture', *Progress in Petrochemical Science*, 5(1). Available at: https://doi.org/10.31031/pps.2023.05.000604.

10. Machine Learning in Materials Science, 2022. YouTube. Available at: https://www.youtube.com/watch?v=WOYVakUWURM&t=98s.

10.APPENDICES

Model 1: Using SiteStatsFingerprint featurizer.

```
[imports and dataset loading]
# Required imports
from matminer.featurizers.structure import SiteStatsFingerprint
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import pandas as pd
from matminer.datasets import load_dataset
from pymatgen.core import Structure
import matplotlib.pyplot as plt
import pickle
# Load the dataset
dataset = load_dataset("matbench_perovskites")
df = pd.DataFrame(dataset)
# Drop rows with missing values
df = df.dropna()
# Check the type of the first element in the 'structure' column
first_element = df['structure'].iloc[0]
print (f"Type of first element in 'structure' column: {type(first_element)}")
# %% [Feature extraction]
from matminer.featurizers.site import OPSiteFingerprint
# Initialize the SiteStatsFingerprint featurizer with OPSiteFingerprint
site_featurizer = OPSiteFingerprint()
structure_featurizer = SiteStatsFingerprint(
    site_featurizer=site_featurizer,
    stats = ('mean', 'std_dev', 'minimum', 'maximum')
```

```
)
# Featurize the 'structure' column with parallel processing
X = structure_featurizer.featurize_dataframe(df, 'structure')
# Save featurized data for future use
with open ('featurized_data.pickle', 'wb') as f:
    pickle.dump(X, f)
# Load featurized data
with open ('featurized_data.pickle', 'rb') as f:
    X = pickle.load(f)
Print ("Featurized data has been loaded successfully.")
[Target variable and train-test split]
y = dataset['e_form']
# Target variable: formation energy
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# %% [Check for missing or non-numeric columns]
# Check if all columns are numerical
print(X_train.dtypes)
import numpy as np
# Identify columns with non-numeric or problematic data
non_numeric_columns = []
for col in X_train.columns:
    if isinstance(X_train[col].iloc[0], (list, np.ndarray)):
        non_numeric_columns.append(col)
    elif not np.issubdtype(X_train[col].dtype, np.number):
        non_numeric_columns.append(col)
# Example of handling array-like columns (you can aggregate them)
for col in non_numeric_columns:
    print (f"Processing column: {col}")
    if isinstance(X_train[col].iloc[0], (list, np.ndarray)):
```

```
        X_train[f'{col}_mean'] = X_train[col].apply(lambda x: np.mean(x) if isinstance(x, (list, np.ndarray)) else x)
        X_train[f'{col}_max'] = X_train[col].apply(lambda x: np.max(x) if isinstance(x, (list, np.ndarray)) else x)
        X_train[f'{col}_min'] = X_train[col].apply(lambda x: np.min(x) if isinstance(x, (list, np.ndarray)) else x)
    # Drop the original array column after aggregation
    X_train = X_train.drop(columns=[col])
# Now your X_train should only contain numeric columns
print(X_train.dtypes)
import numpy as np
# Identify columns with non-numeric or problematic data
non_numeric_columns = []
for col in X.columns:
    # Check if the column contains lists or arrays
    if isinstance(X[col].iloc[0], (list, np.ndarray)):
        non_numeric_columns.append(col)
    elif not np.issubdtype(X[col].dtype, np.number):
        non_numeric_columns.append(col)
# Aggregate array-like columns into numeric features
for col in non_numeric_columns:
    print(f"Processing column: {col}")
    if isinstance(X[col].iloc[0], (list, np.ndarray)):
        # Example: Aggregate using mean, max, and min
   X[f'{col}_mean'] = X[col].apply(lambda x: np.mean(x) if isinstance(x, (list, np.ndarray)) else x)
   X[f'{col}_max'] = X[col].apply(lambda x: np.max(x) if isinstance(x, (list, np.ndarray)) else x)
   X[f'{col}_min'] = X[col].apply(lambda x: np.min(x) if isinstance(x, (list, np.ndarray)) else x)
    # Drop the original non-numeric column
    X = X.drop(columns=[col])
# Verify that all columns are now numeric
print(X.dtypes)
```

```
print ("All columns are now numeric!")
# %% [Cross-validation implementation]
# Define the model (Random Forest Regressor)
model = RandomForestRegressor(random_state=42)
# Initialize K-Fold Cross-Validation (5 splits in this example)
kf = KFold(n_splits=5, shuffle=True, random_state=42)
# Perform cross-validation (using Mean Squared Error as scoring metric)
cv_scores = cross_val_score(model, X, y, cv=kf, scoring='neg_mean_squared_error')  # Negative MSE for sklearn's cross_val_score
# Convert negative MSE to positive
cv_scores = -cv_scores
# Display cross-validation results
print("Cross-Validation MSE Scores:", cv_scores)
print("Mean CV MSE:", cv_scores.mean())
print("Standard Deviation of CV MSE:", cv_scores.std())
# %% [Model training]
# Train the model on the entire dataset
model.fit(X_train, y_train)
# %% [Ensure consistent columns for testing]
# Ensure that X_test has the same columns as X_train
X_test = X_test[X_train.columns]
# %% [Predictions and Evaluation]
# Make predictions on both train and test sets
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)
# %% [Metrics: MAE, RMSE, R²]
# Evaluate the model
print(f"Train MAE: {mean_absolute_error(y_train, y_pred_train)}")
print(f"Test MAE: {mean_absolute_error(y_test, y_pred_test)}")
print(f"Train RMSE: {np.sqrt(mean_squared_error(y_train, y_pred_train))}")
```

```
print(f"Test RMSE: {np.sqrt(mean_squared_error(y_test, y_pred_test))}")
print(f"Train R²: {r2_score(y_train, y_pred_train)}")
print(f"Test R²: {r2_score(y_test, y_pred_test)}")
```

Model 2: Using XRD Calculator featurizer.

```
# Required Imports
import matminer
from pymatgen.analysis.diffraction.xrd import XRDCalculator
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, cross_validate
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score, make_scorer
import pandas as pd
import numpy as np
from matminer.datasets import load_dataset
import matplotlib.pyplot as plt
import seaborn as sns
import pickle
# 1. Load the 'matbench_perovskites' dataset
dataset = load_dataset('matbench_perovskites')
dataset = dataset.dropna()
# 2. Prepare data (features and target)
structures = dataset['structure'] # Perovskite structures in pymatgen Structure format
y = dataset['e_form'] # Target variable: formation energy
# Initialize the XRDCalculator object
xrd_calculator = XRDCalculator()
# Function to extract XRD features: peak intensities
def extract_xrd_features(structure):
    pattern = xrd_calculator.get_pattern(structure)  # Compute XRD pattern
    return pattern.y  # Return the intensities (y-values) of the diffraction peaks
```

```
# Featurize each structure
X = [extract_xrd_features(structure) for structure in structures]
# Save the features DataFrame X to a pickle file
with open('features_X.pkl', 'wb') as f:
    pickle.dump(X, f)
# Load the features DataFrame X from the pickle file
with open('features_X.pkl', 'rb') as f:
    X_loaded = pickle.load(f)
# Convert the list of features into a DataFrame
X = pd.DataFrame(X_loaded)
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize and train the Random Forest model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
# 3. Implement Cross-Validation
# Define the scoring metrics
scoring = {
    'MAE': make_scorer(mean_absolute_error),
    'RMSE': make_scorer(mean_squared_error, squared=False),  # Root Mean Squared Error
    'R2': make_scorer(r2_score)
}
# Perform cross-validation
cv_results = cross_validate(
    rf_model,       # Random Forest Regressor
    X,              # Feature matrix
    y,              # Target variable
    cv=5,           # Number of folds
    scoring=scoring, # Scoring metrics
    return_train_score=True,
```

```
    n_jobs=-1       # Use all available cores
)
# Display cross-validation results
print("Cross-Validation Results:")
for metric in scoring.keys():
    print(f"Mean {metric} (Train): {np.mean(cv_results[f'train_{metric}']):.4f}")
    print(f"Mean {metric} (Test): {np.mean(cv_results[f'test_{metric}']):.4f}")
    print(f"Std {metric} (Test): {np.std(cv_results[f'test_{metric}']):.4f}")
    print()
# Make predictions on the test set
y_pred_train = rf_model.predict(X_train)
y_pred_test = rf_model.predict(X_test)
# Fixed RMSE calculation
def calculate_rmse(y_true, y_pred):
    return mean_squared_error(y_true, y_pred, squared=False)
# Evaluation Metrics
print(f"Train MAE: {mean_absolute_error(y_train, y_pred_train):.4f}")
print(f"Test MAE: {mean_absolute_error(y_test, y_pred_test):.4f}")
print(f"Train RMSE: {calculate_rmse(y_train, y_pred_train):.4f}")
print(f"Test RMSE: {calculate_rmse(y_test, y_pred_test):.4f}")
print(f"Train R²: {r2_score(y_train, y_pred_train):.4f}")
print(f"Test R²: {r2_score(y_test, y_pred_test):.4f}")
```