Bridging Density Functional Theory and Machine Learning: Predicting Formation Energies of Oxide Perovskites

Divya T L [1], Chandrani Chakravorty [2], Shashanka[3], Shashank Hegde[4]

[1,2]Assistant Professor, [3,4]Student
[1,2,3,4]Department of MCA, RV College of Engineering.

Abstract:
Perovskite oxides with the general formula ABO_3 have formed the cornerstone in creating high functional materials because of their great tunability and versatility. Their application spans solar energy, superconductivity, catalysis, and fuel cells. Thermodynamic stability, expressed through formation energy, governs their usefulness. Conventionally, Density Functional Theory (DFT) is used to compute formation energies, but it is computationally expensive. In this work, we propose a machine learning-based prediction of perovskite formation energies using descriptors derived from elemental and structural features. After feature engineering and selection, several models were evaluated. XGBoost achieved the best R^2 score of 0.9586, followed closely by Random Forest. These results demonstrate the potential of ensemble-based machine learning models in accelerating perovskite discovery.

Keywords: Formation Energy, Machine Learning, Random Forest, Oxide Perovskites, XGBoost.

1. Introduction

Based on outstanding performing properties and structural flexibility, Perovskite oxides with a generic chemical formula ABO_3 have received enormous attention in contemporary materials science [1]. Elastic perovskite materials find ap- plications in various technologies such as solid oxide fuel cells, piezoelectric sensors, superconductors, and catalyze Solar energy systems because they exhibit varied electronic, magnetic, and catalytic properties. Composition and structure modification options for these materials has opened new avenues of research in search of advanced variants which leads to heightened performance efficiency. However, exploring the vast compositional space of perovskites is limited by time and resource constraints.

The formation energy of a compound, which is the energy required to form it from its constituent elements in their standard states, dictates the stability and synthesizability of perovskites. A perovskite with a highly negative formation energy will be more stable and easier to synthesize, while one with positive or mildly negative formation energies would need specialized conditions for synthesis or decompose into competing phases. In order to devise accurate models that accelerate the discovery of novel perovskites, it is pivotal first to address the issue of predictive accuracy at the engineering level alongside material selection as formation energy plays a crucial role when designing new materials.

DFT has been the primary method for predicting formation energies due to its reasonable accuracy and computational feasibility, despite its quantum mechanical limitations [2]. DFT has greatly contributed to our understanding of perovskite stability, but its shortcomings—such as how it treats electron correlations, the large-scale computational cost for extensive screening, among others large scale computational costs for extensive screening—have forced researchers to look for other methods. Over the last few years machine learning (ML) has surfaced as a potent alternate solution since it can efficiently predict new compositions from existing structure property relationships databases. Achieving a blend between accuracy and computational costs through DFT and ML makes it possible for researchers conduct high-throughput screening of perovskite stability.

2. Data Limitations

Continuously estimating oxide perovskite formation energy still presents multi-faceted challenges to computational materials science due to data gaps, variability of methodologies, and the complex nature underlying the stability of perovskites. One of the significant limitations is that there is no harmonized, broad dataset. This is because existing [computational databases] such as the Materials Project, OQMD, or AFLOW contain varying parameter values for required parameters which make direct data merging difficult [3]. These discrepancies arise due to other forms of exchange-correlation functionals (GGA-PBE, PBEsol, SCAN) different approaches due to Hubbard U correction methods for transition metals or technical set default parameters such as pseudopotentials, energy cutoffs as well as relaxation procedures. Such methodological discrepancies result in systematic errors that significantly influence calculated formation energies producing tens of meV/atom differences based on the computation method.

Such gaps can be especially impactful in the case of machine learning methods aimed at materials discovery and innovation. Predictive models trained on mixed datasets with outcomes from various computational approaches are sure to create these artificial gaps rather than real structure-property correlations that are essential. This issue is most striking in comparisons between standard DFT functionals like GGA- PBE and more sophisticated ones, including hybrid functionals HSE06, which can yield significantly different results. Because of such challenges, we need to pay attention to specially crafted, methodologically uniform datasets aimed at developing predictive models for perovskite formation energies. Also, this highlights the lack of universally accepted computational standards within the community that would facilitate depend- able data integration and model construction.

In this study used data from the Materials Project (MP) to ensure internal consistency in the computational parameters used for formation energy calculations. While this approach avoids the pitfalls of merging disparate datasets with varying exchange-correlation functionals (e.g., PBE vs. SCAN), Hubbard U corrections, or relaxation criteria, it introduces inherent limitations. The MP dataset relies on GGA-PBE, which systematically underestimates formation energies for strongly correlated systems (e.g., Co-, Ni-, or Mn-based perovskites) and may overrepresent computationally stable phases at the expense of experimentally realized but metastable compositions. These biases are propagated into machine learning models trained on such data, as no universal corrective scheme exists for GGA-PBE's errors. However, this trade-off is justified by the critical need for methodological uniformity when developing predictive models.

Although it is acknowledged that the predictions of this work will inherit the limitations of DFT, the main goal is to train an accurate machine learning model for formation energy prediction. By concentrating on a single, carefully selected database, we reduce the limitation caused by combining disparate data sources and offer a repeatable starting point for further research.

3. Methodology

Having demonstrated the difficulties in perovskite formation energy prediction based on DFT and the promise of machine learning (ML) to bridge the gap, next step is to outline predictive modeling framework. This work emphasizes a methodical strategy for preprocessing, feature engineering, and model construction based on the Materials Project (MP) dataset for ensuring consistency.

In order to train a good machine learning model to predict the formation energy of oxide perovskites [4], we used a dataset of 2,503 ABO_3-type perovskite compounds from the Materials Project (MP) database. The dataset preprocessing was the most important step to address data quality [5] and model trustworthiness and was made up of three steps: handling null values, encoding categorical data, and standardizing features.

The initial preprocessing step was to deal with missing values in the data, typically occurring due to incomplete simulations or data source inconsistencies. Most machine learning models can't function effectively with missing values, so we adopted a two-stage approach. We deleted records where critical features such as formation energy or elemental properties were missing, as these are required for correct predictions. For less severe gaps, like intermittent missing ionic radii values, median imputation was employed to replace them. This was crucial to make sure the data that entered the prediction models was complete and valid, lowering the chances of bias or false outcomes.

In the next step, categorical variables were transformed into numerical form. This involved features such as space group symmetries (e.g., Pm-3m, Pnma) and elemental occupations at A and B sites [6]. As the majority of machine learning algorithms take numerical inputs, we employed one-hot encoding for features with fewer unique categories, like space groups. In the case of features with a built-in order—such as oxidation states—we utilized ordinal encoding. Getting this step correct was crucial, since categorical labels can have a significant impact on the structural stability of perovskites. Incorrect encoding could result in deceptively misleading numerical patterns that would harm the way the model learns from the data.

As a last step in preprocessing, we normalized all the numerical features so that they were on the same scale, converting each feature to have a mean of zero and a standard deviation of one [7]. Normalization was especially crucial since features like ionic radii (which are measured in Angstroms) and formation energies (which are measured in eV/atom) naturally occur on different scales. Without standardization, machine learning algorithms, particularly distance metric- or gradient-based optimization algorithms, may disproportionately emphasize high-magnitude features, which would result in poor model performance. Standardization also accelerated convergence rates for neural networks and support vector machines, leading to improved training efficiency.

These preprocessing operations collectively improved the quality of the dataset, so that the following machine learning model was able to learn meaningful structure-property relationships instead of data inconsistency artifacts. Through systematic handling of missing values, correct encoding of categorical variables, and scaling of feature values, we created a solid basis for precise formation energy prediction. Feature selection and model training are the subsequent steps of this work, using this fine-tuned dataset to realize high predictive accuracy.

The choice of the right features is an important aspect of creating good machine learning models for perovskite formation energy prediction. To make sure that we chose descriptors that are statistically relevant and physically valid, we used two approaches: machine learning-based importance ranking and statistical correlation analysis. Instead of using them together upfront, we used each one separately. This allowed us to look at the properties from both the mathematical and physical points of view and made our choice process stronger.

Performed correlation analysis to identify descriptors with absolute correlation coefficients greater than 0.25 relative to the target formation energy and employed Random Forest regression specifically for feature importance analysis, which provided additional insight beyond simple correlation metrics. On the basis of the output of these two methods identified the following descriptors as of highest importance:

- Energy above hull (quantifying thermodynamic stability relative to alternative phases)
- Band gap (quantifying electronic structure properties)
- Energy per atom (total system energy descriptor)
- Valence band maximum (highest occupied electronic states descriptor)
- Conduction band minimum (lowest unoccupied electronic states descriptor)

The chosen subset of descriptors provides the best com- promise between model simplicity and predictivity, without compromising on the danger of overfitting characteristics of too many parameters included. These physically relevant parameters will provide the basis for constructing predictive models.

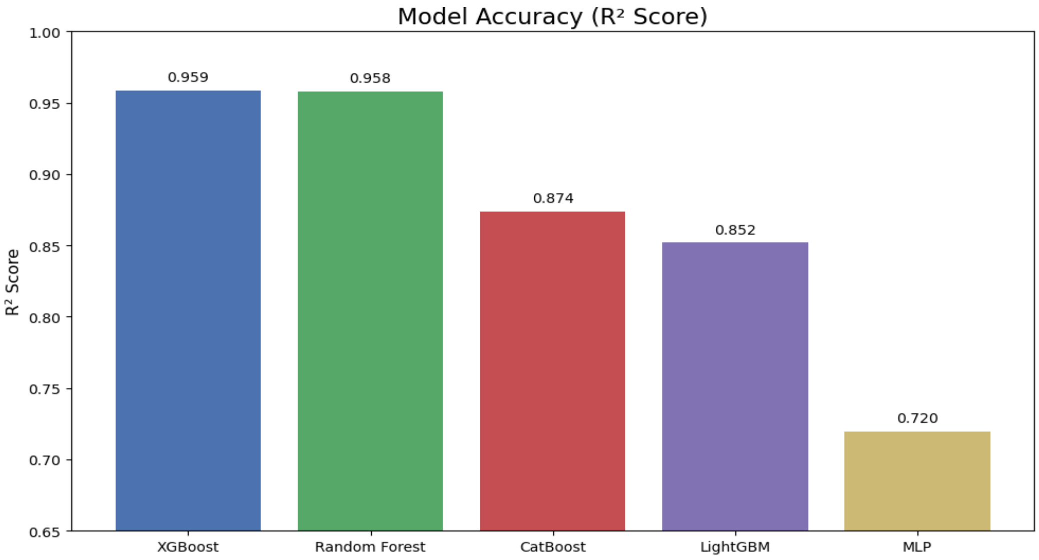
With the chosen descriptors constituting a concise yet detailed feature set, several regression algorithms were used to evaluate their performance in the prediction of the formation energy of perovskite compounds. Some of the machine learning models used were Random Forest Regressor, Light Gradient Boosting Machine (LightGBM), CatBoost, Multi- layer Perceptron (MLP), and Extreme Gradient Boosting (XG- Boost). These were selected due to their varied architectures and established effectiveness in addressing regression tasks with intricate, high dimensional datasets.

For consistency in the evaluation, all the models were trained on the same dataset with the same pre-processing techniques. The dataset was split into training and testing subsets in 80:20 proportions and five-fold cross-validation was employed to further test the generalization performance.

4. Results

A comparison of R^2 scores—indicating how well every model learned from the data at hand—is depicted in Figure 1.

Figure. 1. Model Accuracy Based on R2 Scores



As can be seen in the figure, XGBoost performed the best, with an R^2 value of 0.9586, slightly better than the Random Forest Regressor, which had 0.9577. This suggests that both of the ensemble-based techniques were very good at modelling complex, non-linear relationships between the descriptors chosen and the target property.

The second best accuracy was for CatBoost at $R^2 = 0.8737$, followed by LightGBM at $R^2 = 0.8518$. Although slightly less precise than Random Forest and XGBoost, these models also performed well and verified that gradient-boosted decision trees are appropriate for this field.

At the lower end, the Multilayer Perceptron model produced a R^2 of 0.7198, reflecting relatively poor capacity to fit training data. This result may be a consequence of the intrinsic difficulty in neural networks and their hyperparameter sensitivity, particularly when trained on highly structured data of small size or variance.

The stark contrast between the best-performing models (Random Forest and XGBoost) and the neural network model highlights the significance of choosing appropriate algorithms specific to the domain of data. Within materials informatics, where feature space is organized and the volume of data is comparatively small, ensemble tree-based models often exhibit optimal performance because of their capacity to capture interaction, address feature heterogeneity, and provide robustness against overfitting.

Based on this comparison, XGBoost was chosen to further develop and test, due to its high learning ability and consistent accuracy among folds. These results confirm that the chosen descriptors are not just physically significant but also statistically robust enough to support high-performing predictive models.

5. Conclusion

In this research, we have investigated a machine learning- based data-driven method to estimate the formation energies of oxide perovskites (ABO_3), offering a computationally effective alternative to established quantum mechanical approaches like density functional theory (DFT). By using a carefully compiled dataset from the Materials Project (MP), we sought to ensure computational homogeneity and steer clear of mixed or hybrid artifacts present in combining multiple datasets. In spite of recognized constraints related to GGA-PBE functionals, the exclusive use of MP data ensured a uniformity of method, an indispensable parameter for creating robust machine learning models.

The essence of this work is to prove the capability of machine learning models, specifically tree based ensemble methods, in precisely estimating formation energies with a physically interpretable and compact set of descriptors. The statistical correlation and model-based importance ranking guided feature selection process enabled to narrow down important descriptors such as energy above hull, band gap, valence band maximum, and conduction band minimum. These were not only selected based on statistical significance but also physical importance to characterize material stability and electronic structure.

The following multiple regression models were tried and compared against each other for their capacity to learn intractable structure-property relationships: Random Forest Regressor, LightGBM, CatBoost, XGBoost, and Multilayer Perceptron (MLP). Of the above, XGBoost had the most consistent training accuracy with an R^2 value of 0.9586, followed very closely by Random Forest. These findings confirmed the success of ensemble models based on trees in the case of perovskite formation energy prediction, particularly when dealing with structured data of moderate variance. Neural networks, though strong in larger unstructured spaces, were behind in performance as a result of size and complexity issues with the data.

One of the strongest conclusions to be drawn from this research is that machine learning can be an inexpensive in terms of computation, scalable, and relatively accurate surrogate for DFT in certain materials science applications. By constructing models capable of good generalization over huge compositional spaces, this research opens the door to high- throughput screening of perovskite compounds—a critical step in fast-tracking the discovery of new functional materials for energy, electronics, and catalysis applications.

Nevertheless, this research also stresses that ML is only as good as the data and domain knowledge behind it. The recognized biases of the Materials Project dataset and the potential weaknesses in the GGA-PBE functional are carried forward in our models, pointing to the need for ongoing refinement of more complete and error-corrected databases. Moreover, extending the dataset to comprise experimentally validated or hybrid DFT-calibrated data—without compromising on consistency— may further enhance the generalization and predictive capacity of the model.

Finally, this paper presents a well-organized, replicable approach to machine learning prediction of perovskite formation energies, with explicit evidence of its utility and accuracy in materials discovery. Directions for future work will involve incorporating more advanced descriptors (e.g., structural motifs, phonon modes), generalizing the model to include non-oxide perovskites, and investigating transfer learning methods to overcome domain adaptation issues. With such breakthroughs, we see machine learning as an innate part of current computational materials science, speeding up the path from atomic structure to useful applications.

6. Future Enhancements

More advanced and physically descriptive descriptors than the existing ones can be used in future research. Additional features based on the crystal symmetry parameters, Bader charges, partial density of states, as well as the phonon dispersion characteristics, can improve the model in terms of the quality of the description of the nuance in the structure property relationship. The addition of geometric and topo- logical data (e.g. Voronoi tessellation features, coordination environment, or structural motifs) have the potential to enable models better interpret spatial interactions and lattice dynamics and lead to increased accuracy and generalizability. The method can fill the gap between local atomic surroundings and global stability characteristics.

Although this work focused on oxide-based ABO3 perovskites, extension in future could include non oxide perovskites (e.g. halide, nitride, or sulfide perovskites), double perovskites (A2BB’O6), and hybrid organic-inorganic compounds. That would enlarge the chemical space and practical applicability of the model much more, especially as it would be applied in photovoltaics and thermoelectric. In order to do this domain adaptation or multi-task learning frameworks can be used which allows the re-use of learned representations across material families. Besides, the transfer learning experience can be used as a mitigation of data scarcity problems in consideration of less common types of compounds.

An interesting improvement would be the addition of formation energies that have been verified by experiment, or computed with still higher-accuracy methods, like hybrid functionals or many-body perturbation theory. Although it is hard to make data consistent, it may be possible to combine such data in a thoughtful way, possibly by normalizing that data by domain, or by using an ensemble-model basis, to alleviate bias caused by GGA-PBE (Perdew-Burke-Ernzerhof (PBE) functional within the Generalized Gradient Approximation).

REFERENCES:

[1] K. Monareng, D. Tshwane, P. Ntoahae, and R. Maphanga, “Enhanced machine learning approaches for predicting formation energy and tolerance factor in perovskite oxide materials,” in Proc. 68th Annu. Conf. South African Inst. Phys. (SAIP), Rhodes University, South Africa, 2024.

[2] S. P. Shaji and W. Tress, “Data-driven analysis of hysteresis and stability in perovskite solar cells using machine learning: Can machine learning help to extract hidden correlations from the perovskite

database? – A case study on hysteresis and stability,” Energy and AI, vol. 20, p. 100503, 2025. doi: 10.1016/j.egyai.2025.100503

[3] S. Touati, A. Benghia, Z. Hebboul, I. K. Lefkaier, M. B. Kanoun, and S. Goumri-Said, “Machine learning models for efficient property prediction of ABX3 materials: A high-throughput approach,” ACS Omega, vol. 9, no. 48, pp. 47519–47531, Dec. 2024. doi: 10.1021/acsomega.4c06139

[4] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, and A. A. Lee, “Rapid discovery of stable materials by coordinate-free coarse graining,” *Science Advances*, vol. 8, no. 30, p. eabn4117, 2022. [Online]. Available: https://doi.org/10.1126/sciadv.abn4117

[5] H. A. H. Mahmoud, “Computerized prediction of perovskite performance using deep learning,” Electronics, vol. 11, no. 22, p. 3759, 2022. doi: 10.3390/electronics11223759

[6] G. S. Thoppil and A. Alankar, “Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning,” Computational Materials Science, vol. 211, p. 111506, 2022. doi: 10.1016/j.commatsci.2022.111506

[7] X. Li, J. Williams, C. Swanson, and T. Berg, “A machine learning approach to predictive maintenance: Remaining useful life and motor fault analysis,” Computers & Industrial Engineering, vol. 206, p. 111222, 2025. [Online]. Available: https://doi.org/10.1016/j.cie.2025.111222