

# BENCHMARKING TREE-BASED MACHINE LEARNING MODELS FOR FORMATION ENERGY PREDICTION IN PEROVSKITES

A. S. Lihitkar

Department of Physics,  
Indira Gandhi Kala Mahavidyalaya,  
Ralegaon, Dist. Yavatmal  
Email : [amolpralay@gmail.com](mailto:amolpralay@gmail.com)

Crossref DOI - <https://doi.org/10.63665/rh.v7i2.60>

## Abstract :

*Machine learning (ML) has become an important tool for accelerating materials discovery by enabling rapid prediction of material properties from existing datasets. Among various material classes, perovskites have attracted significant attention due to their applications in photovoltaics, catalysis, and electronic devices. In this work, four tree-based machine learning models are benchmarked for the prediction of formation energies of perovskite materials. These are Random Forest, Gradient Boosting, XGBoost, and LightGBM. The Matbench perovskite dataset, consisting of 18,928 crystal structures with density functional theory (DFT)-calculated formation energies, was used. Composition-based features were generated using Magpie descriptors, while structure-based features were obtained from density and symmetry properties. Models were trained using an 80/20 train-test split and evaluated with five-fold cross-validation. Among the tested models, XGBoost achieved the best performance with a test mean absolute error (MAE) of 0.227 eV/atom and an  $R^2$  value of 0.79. The results indicate that boosting-based tree models outperform bagging-based approaches and demonstrate the importance of structural descriptors for accurate prediction of perovskite formation energies.*

**Keywords** : Perovskites; Machine learning; Formation energy; Tree-based models; XGBoost; Materials informatics

## Introduction :

The development of functional materials is central to advances in energy, electronics, and catalysis. Traditionally, materials discovery has relied on experimental synthesis and computational methods such as density functional theory (DFT). While DFT provides reliable predictions of material properties, its computational cost limits its applicability to large-scale materials screening [1].

Machine learning (ML) has emerged as a promising alternative for predicting material properties by learning patterns from existing datasets [2]. ML models can significantly accelerate materials discovery by providing near-instantaneous predictions once trained. This approach has been successfully applied to predict formation energies, band gaps, elastic



properties, and other materials characteristics [2,3].

Perovskite materials, typically represented by the general formula  $ABX_3$ , have attracted extensive research interest due to their applications in solar cells, ferroelectrics, catalysts, and sensors [4]. The vast compositional and structural diversity of perovskites makes them an ideal candidate for machine learning-based screening approaches.

Among various ML algorithms, tree-based ensemble methods are widely used in materials informatics because of their robustness, ability to handle nonlinear relationships, and minimal preprocessing requirements [2]. Random Forest and Gradient Boosting models have shown strong performance in predicting material properties [2,5]. More recently, optimized boosting algorithms such as XGBoost and LightGBM have gained popularity due to their improved accuracy and computational efficiency [6,7].

Despite the widespread use of tree-based models, systematic comparisons of these algorithms for perovskite formation energy prediction remain limited. Therefore, the objective of this study is to benchmark four widely used tree-based models—Random Forest, Gradient Boosting, XGBoost, and LightGBM—using the Matbench perovskite dataset [3].

## **Dataset and Feature Engineering :**

### **1. Dataset :**

This analysis uses the Matbench perovskite dataset. This dataset is a component of the Matbench benchmark suite, which is intended for the assessment of materials science machine learning models [3]. The dataset includes 18,928 structures made of perovskites. DFT was used to calculate formation energies and crystallographic structure information. The target variable was formation energy per atom.

### **2. Composition-Based Features :**

Composition-based features were generated using the Magpie descriptor set implemented in the Matminer library [2]. These features capture statistical properties of the constituent elements, including: Atomic number, Electronegativity, Atomic radius, Valence electron counts and Periodic table statistics. Magpie features have been widely used in materials informatics and have demonstrated strong predictive capability [2].

### **3. Structure-Based Features :**

To incorporate crystallographic information, structure-based features were extracted using the Matminer library [2]. Two categories of structural descriptors were considered: density-related features and symmetry-related features. The density features included properties such as bulk density, volume per atom, and packing efficiency, which capture the geometric and spatial characteristics of the crystal structure. In addition, symmetry features were incorporated to describe the crystallographic arrangement of atoms, including the crystal system and space group number. Since certain symmetry descriptors, such as the crystal system, are categorical in nature, they were converted into numerical form using one-hot encoding to ensure compatibility with the machine learning models.



#### **4. Feature Preprocessing :**

The data preprocessing workflow involved several critical stages to ensure a robust feature set for model training. Initially, composition and structural features were combined into a unified feature matrix. Categorical variables were then transformed using **one-hot encoding** to facilitate compatibility with tree-based algorithms. To refine the feature space, columns containing **missing values (NaN)** and **constant features**—which offer no predictive variance—were systematically removed. Following these refinement steps, the final processed dataset comprised **18,928 samples** characterized by **129 numerical features**, providing a high-dimensional foundation for subsequent energy prediction modeling

#### **Machine Learning Models :**

In this study, four tree-based machine learning models were selected and benchmarked for predicting the formation energies of perovskite materials. Tree-based methods are widely used in materials informatics because they can capture nonlinear relationships, handle complex feature interactions, and require minimal data preprocessing. These models also provide strong predictive performance across a wide range of materials datasets.

##### **1. Random Forest :**

Random Forest is an ensemble learning method that constructs a large number of decision trees during training. Each tree is built using bootstrap sampling of the training data and a random subset of features at each split, which introduces diversity among the trees. The final prediction is obtained by averaging the outputs of all trees, which reduces overfitting and improves generalization [5].

##### **2. Gradient Boosting :**

Gradient Boosting is a sequential ensemble technique in which trees are built one after another. Each new tree is trained to correct the residual errors made by the previous trees. By gradually minimizing the loss function, the model improves its predictive accuracy. This approach allows the model to capture complex nonlinear relationships between features and target properties [8].

##### **3. XGBoost :**

XGBoost is an optimized implementation of gradient boosting designed for speed and performance. It introduces regularization terms to prevent overfitting and uses parallel computation for faster training. Additionally, XGBoost employs efficient tree construction and advanced optimization techniques, making it one of the most widely used algorithms in machine learning competitions and materials informatics studies [6].

##### **4. LightGBM :**

LightGBM is another gradient boosting framework that focuses on computational efficiency and scalability. It uses histogram-based feature binning and a leaf-wise tree growth



strategy, which allows it to achieve high accuracy with reduced training time. This makes LightGBM particularly suitable for large datasets with many features, such as those commonly encountered in materials science applications [7].

## Model Training and Evaluation :

### 1. Train–Test Split :

The dataset was divided into two subsets to evaluate model performance: 80% of the data was used for training the machine learning models, while the remaining 20% was reserved as an independent test set. The training set was used to learn the relationships between the input features and the formation energy, whereas the test set was used to assess the predictive performance of the models on unseen data. This split ensures a fair evaluation and helps prevent overfitting.

### 2. Cross-Validation :

The robustness of the four tree-based models was tested using five-fold cross-validation. In this method, the training data were divided into five equal parts. Each part was used once for validation while the remaining parts were used for training the model. This process helped in checking how well the models perform on unseen data and reduced the chances of overfitting. The results showed that the ensemble methods and optimization techniques used in these models improved the prediction accuracy. Overall, the cross-validation confirmed that these models can successfully learn the complex and nonlinear patterns present in perovskite datasets [9].

### 3. Performance Metrics :

Two evaluation metrics were used to measure the performance of the models: Mean Absolute Error (MAE) and the coefficient of determination ( $R^2$ ). The MAE represents the average absolute difference between the predicted and actual formation energy values, and is given by

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

where  $y_i$  is the actual value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of samples.

The  $R^2$  score indicates how well the model explains the variation in the data. It is calculated using

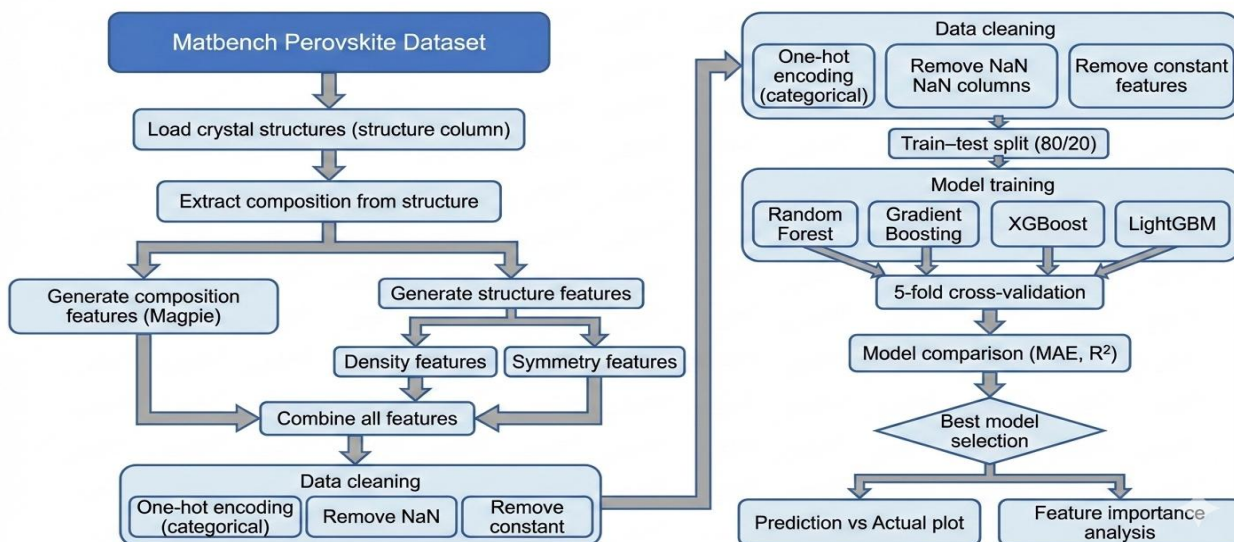
$$\text{Coefficient of determination (R}^2\text{)} = 1 - \frac{\sum |y_i - \hat{y}_i|}{\sum |y_i - \bar{y}|}$$

where  $\bar{y}$  is the mean of the actual values. A lower MAE and a higher  $R^2$  indicate better model performance [10].

### 4. Machine Learning Workflow :



**Flowchart of the Machine Learning Pipeline Used in This Study.**



**Figure 1** Flowchart of the machine learning pipeline used in this study.

**Results and Discussion :**

**1. Cross-Validation Performance :**

The cross-validation results of the four tree-based models are presented in Table 1. The results show that the boosting-based models perform better than the Random Forest model. Among all the models, XGBoost achieved the lowest cross-validation error, indicating its superior predictive capability for the dataset.

**Table 1 :** Cross-validation performance of tree-based models.

Model	CV (eV/atom)	MAE
Random Forest	0.274	
Gradient Boosting	0.265	
LightGBM	0.243	
<b>XGBoost</b>	<b>0.237</b>	

**Table 2 :** Test set performance comparison.

Model	Test (eV/atom)	MAE	R <sup>2</sup>
Random Forest	0.267		0.75
Gradient Boosting	0.265		0.77
LightGBM	0.241		0.80
<b>XGBoost</b>	<b>0.227</b>		<b>0.79</b>

**2. Test Set Performance :**

The performance of the models on the independent test set is shown in Table 2. XGBoost achieved the lowest MAE among all models, while LightGBM showed a slightly higher R<sup>2</sup> value. Both boosting-based models performed better than Random Forest and Gradient Boosting.

**3. Model Comparison :**

The ranking of models based on test MAE is :

$$XGBoost > LightGBM > Gradient Boosting > Random Forest$$



Boosting-based models consistently outperformed the bagging-based Random Forest model. This trend is consistent with previous studies showing the superior performance of gradient boosting methods in materials property prediction [2,6].

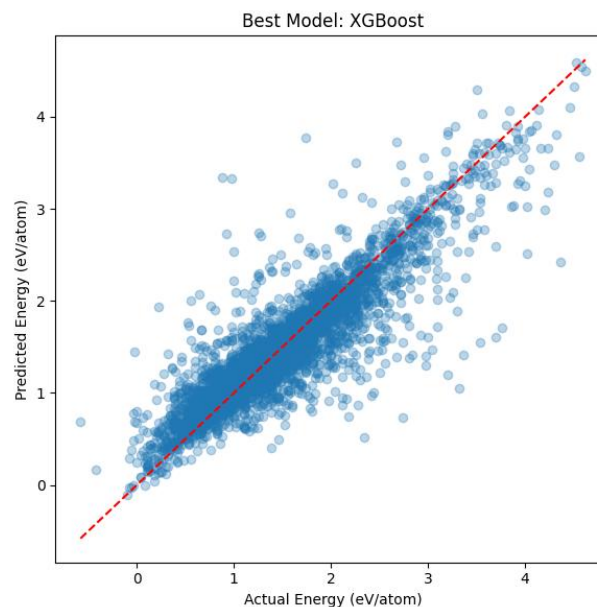
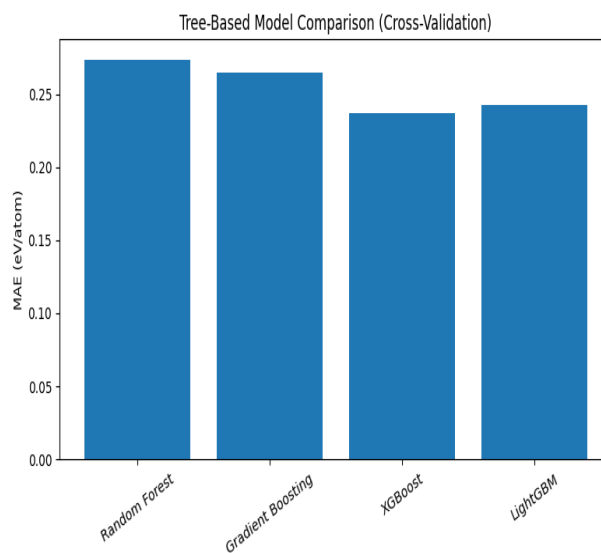
#### 4. Effect of Structural Features :

Initial experiments using only composition-based features produced MAE values in the range of 0.45–0.58 eV/atom. The data is not shown in this paper. After incorporating structure-based descriptors, the MAE decreased to approximately 0.23 eV/atom.

This represents nearly a **50% reduction in prediction error**, highlighting the critical importance of structural information in predicting perovskite stability. Similar observations have been reported in previous materials informatics studies [2,3].

#### 5. Prediction Performance :

The figure 2 compares the cross-validation mean absolute error (MAE) of four tree-based machine learning models for perovskite formation energy prediction. Boosting-based models, particularly XGBoost and LightGBM, show lower MAE values than Random Forest, indicating better predictive performance. The plot in figure 3 shows a strong correlation between predicted and actual values, indicating reliable model performance for the best performing model i.e. XGBoost.



**Figure 2:** Comparison of five-fold cross-validation mean absolute error (MAE) for different tree-based machine learning models

**Figure 3:** Scatter plot of predicted versus actual formation energies for the best-performing model (XGBoost).

#### Conclusion :

In this study, four tree-based machine learning models were benchmarked on the Matbench dataset to predict formation energies of perovskite materials. Boosting-based models outperformed random forest, with XGBoost achieving the highest performance (MAE



= 0.227 eV/atom,  $R^2 = 0.79$ ). Incorporating structure-based features reduced prediction error by approximately 50%. These results highlight the efficacy of tree-based models for rapid perovskite formation energy prediction and underscore the value of structural descriptors in materials informatics.

However, limitations persist, including reliance on dataset-specific features and limited generalizability to diverse perovskite compositions beyond Matbench. Future work could explore hybrid models integrating graph neural networks with tree-based ensembles, incorporate larger multi-fidelity datasets, and validate predictions via high-throughput experiments to enhance robustness and transferability.

### Acknowledgments :

The author acknowledges the developers of the Matminer and Matbench projects for providing open-source tools and datasets. The author acknowledges the use of ChatGPT (OpenAI) and Google Gemini for language assistance and generation of illustrative figures. All scientific analysis, results, and conclusions were performed and verified by the author.

### References :

- Kohn, W., & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A), A1133.
- Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1), 16028.
- Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: The Matbench test set and automated leaderboard. *npj Comput. Mater.* **2020**, 6, 138.
- Kojima, A.; Teshima, K.; Shirai, Y.; Miyasaka, T. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* **2009**, 131, 6050–6051.
- Breiman, L. Random forests. *Mach. Learn.* **2001**, 45, 5–32.
- Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785–794.
- Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* **2017**, 3146–3154.
- Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 29, 1189–1232.
- Valsalakumar, S., Bhandari, S., Roy, A., Mallick, T. K., Hinshelwood, J., & Sundaram, S. (2024). Machine learning driven performance for hole transport layer free carbon-based perovskite solar cells. *Npj Computational Materials*, 10(1). <https://doi.org/10.1038/s41524-024-01383-7>

