JOURNAL OF MACHINE LEARNING AND APPLICATIONS
Jan-Jun 2025, VOL. 1, ISSUE 1, pp. 20-26
https://doi.org/10.61577/jmla.2025.100004

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

Comparative evaluation of ensemble machine learning models for predicting band gaps of double perovskites

M. Noman Nasir[1], Farhan Sarwar[1], Muazzam Ali[2] , Ukasha Ijaz[1], Tehmina Mukhtar[1], Kainat Fatima[3], Junaid Zaidi[1], Muzammil[4]

[1]Department of Basic Sciences, Superior University, Lahore, Pakistan
[2]Department of Computer Science and Information Technology, The Superior University, Lahore, Pakistan
[3]Department of Computer Science, University of South Aisa, Lahore, Pakistan
[4]Department of Mathematics, Ripha International University, Lahore, Pakistan

ABSTRACT

For the discovery of promising new materials for energy and optoelectronic applications, it is of utmost importance to predict the band gaps of double perovskites with high accuracy. Standard Density Functional Theory (DFT) methods, although trustworthy, tend to be computationally expensive compared to big scale screenings. In order to circumvent these limitations, we develop a Gradient Boosting Regression (GBR)-based machine learning (ML) framework strengthened by polynomial feature expansion, standardized preprocessing and rigorous hyperparameter optimization. It has 39 descriptors fed from compositions, such as electronegativity, ionic radii, oxidation states, and orbital energies, giving a dataset of 4,121 double perovskite compounds. Separation of data into training and test sets was done randomly (80:20) and evaluation of model performance was done using more than one metric. The models for GBR yielded 0.9556 (adjusted R^2) 0.0988 eV (MAE), 0.2180 eV (RMSE), 0.9574 (EV), and 4.8393 (RPD) respectively, which indicates that this model was high predictive accuracy and reliability. Compared with earlier works which used Support Vector Regression (SVR), Random Forest and XGBoost, the proposed framework achieves far better performance with physical interpretability. This enables a reproducible data-driven approach between trial and error towards near-DFT-level predictions, leading to both high-throughput screening and rational design of double perovskites.

KEY WORDS

Ensemble machine learning; Support vector regression; Electronic bandgap; XGBoost; LightGBM

ARTICLE HISTORY

Received 29 April 2025;
Revised 20 May 2025;
Accepted 27 May 2025

Introduction

In recent years, double perovskites with general chemical formularies of $A_2BB'X_6$ or $AA'BB'O_6$, have attracted a lot of interests owning to their structural diversity, thermal stability, and tunable optoelectronic properties. Such properties make these materials highly attractive for use in solar cells, photodetectors, thermoelectric, and light-emitting devices [1,2]. The electronic bandgap is a major factor controlling their performance as the bandgap determines optical absorption, charge carrier dynamics and ultimately device efficiency [3,4]. A major hurdle for computational materials science and experimental design is the ability to reliably predict the bandgap over large compositional spaces [5]. Conventional computational methods can provide a decent estimate for the bandgap, such as DFT with different functionals like PBE, HSE06 or GW. Despite this, these approaches are not applicable for the screening of thousands of candidate compositions, due to the computation cost a high precision takes in these methods [6]. Modelling large datasets using high-fidelity DFT calculations requires significant computational resources and time, limiting their use in rapid materials discovery pipelines [7,8]. Similarly, some DFT methods give systematically low bandgaps, necessitating post-processing corrections or hybrid functionals that further raise computing costs [9].

On the other hand, ML based predictive models have proven to be a faster and more cost-effective approach. Existing datasets can be used to train ML models to learn complex nonlinear relationships between elemental descriptors and electronic properties, and once trained, bandgaps can be predicted at a speed and accuracy somewhere between DFT and [10,11]. These methods are well suited for high-throughput screening, where thousands of compositions must be rapidly evaluated to find a few promising candidates [12].

ML has been successfully applied for bandgap prediction, as evidenced by several studies which demonstrate its feasibility [13]. A features-based regression model for oxide double perovskites ($AA'BB'O_6$), using descriptors including electronegativity and lowest occupied Kohn–Sham energy levels on a training set of 1306 compounds, achieving near-DFT accuracy. (2016) Gao et al. for lead-free halide perovskites ($A_2BB'X_6$), XGBoost regression has been previously implemented giving better prediction performance than SVR and Artificial Neural Networks (ANN) for 745 inorganic perovskites. The hierarchical Random Forest system applied to oxide perovskites has converted ~5 million candidate compositions into binary class-words and resulted in 310 double perovskites with the confidence of their stability >90%. An ensemble-learning method using Random Forest, XGBoost, and LightGBM requires considerably less computational cost compared to DFT but retains similar precision.

Nonetheless, a number of limitations in the extant literature remain even in light of these advances. Previous studies tend to be feature redundant, weak interpretability, and lack of cross-domain generalization [13,15]. Furthermore, there is also a lack of standardization in units, splitting protocols as well as

***Correspondence:** Mr. Farhan Sarwar, Department of Basic Sciences, Superior University, Lahore, Pakistan, e-mail: su92-mpmmw-f24-003@superior.edu.pk

external validation which limits reproducibility and reliability of results [16,17].

To overcome these issues, we propose a GBR-based predictive framework for double-perovskites. The model integrates:

Literature Review

Over the past decade, we have seen tremendous advances in the prediction and optimization of electronic bandgaps in double perovskites, particularly due to the integration of ML with DFT [18,19]. Initial efforts concentrated exclusively on oxide perovskites, followed by research on halide and lead-free systems, driven by the demand for environmentally benign and photovoltaic systems [20,21]. Feature-engineered regression models applied to predict bandgaps in oxide double perovskites ($AA'BB'O_6$) [5] With over 1 million possible descriptors calculated from a dataset of 1,306 DFT-compounds derived from the Computational Materials Repository, they identified elemental properties of interest, including electronegativity, ionic radii, and lowest occupied Kohn–Sham orbital energies as key features. We note here the results of their model which was near-DFT accuracy $R^2 \approx 0.900$, proving that judicious descriptor engineering could be capable of capturing important physics while still being computationally efficient [22]. This work serves as a basis for using physics-informed ML in the field of materials informatics. Moving into halide perovskites, Gao et al. For instance, Alenazi et al. utilized XGBoost regression to predict bandgaps of 745 lead-free double perovskites ($A_2BB'X_6$). Compositional descriptors based on those related to atomic number, oxidation state, electronegativity, and ionic radii were utilized in the study. The model gained extreme success when compared with SVR and ANN, yielding an RMSE ≈0.350 eV [23,24]. The candidate structures were screened to identify promising photovoltaic materials and led to the identification of K_2NaInI_6 (1.46 eV) and Na_2MgMnI_6 (1.89 eV, both verified with DFT and ab initio molecular dynamics) after screening >5,000 structures. In this study, they demonstrate the power of hybrid ML–DFT strategies for accelerated materials discovery [25,26].

A hierarchical Random Forest structure based on a dataset of more than 5,000 DFT-generated entries spanning the periodic table and 68 elements, aligned with oxide perovskites. The approach was a two-stage pipeline, in which a classification model was first applied to narrow down candidates to wide-gap materials (Eg ≥ 0.5 eV), followed by regression to provide accurate Eg predictions [27]. This framework discovered 310 double perovskites that are stable and have a confidence level above 0.90, while also offering the design maps that relate elemental substitution, to trends in the bandgap [28,29]. The work presented here showed how combined hierarchical and multi-step ML pipelines may provide for large-scale screening while allowing interpretable structure property relationships to be useful for validation with respect to chemical intuition. More recently, a high-throughput ensemble-learning model based on Random Forest, XGBoost, and LightGBM was used to predict the bandgaps of 1,306 double perovskites. The ensemble method yielded bayesian optimization for hyperparameter tuning, obtained near-DFT accuracy at a lower cost [30]. The analysis of features importance indicated that the one order of descriptors, such as electronegativity difference, ionic radius and orbital energies had dominated the prediction [31,32]. They also highlighted how ensemble learning can enhance the accuracy, reliability and explainability of predictions.

There are trends across these studies

Electronegativity, ionic radius, oxidation state, and orbital energies are all consistent dominants when it comes to descriptor importance.

On the methodological side, it evolved from single-model regression to hybrid ML–DFT screening to hierarchical multi-model frameworks and ensemble learning approaches.

This comes with limitations such as non-cross-domain validation, train/test splitting that may not reflect the real-world problems, and under-reporting of uncertainty metrics, underlining that more reproducible and physical interpretable ML pipelines are needed.

The following literature foundation shows that a pragmatically engineered feature, an ensemble-based ML framework, an adequate validation, and interpretability, makes a ML framework with high predictive accuracy, while maintaining computational efficiency. This work expands on those observations as it improves preprocessing, polynomial feature expansion and evaluation of a GBR model to address the limitations seen in past work.

Methodology

Overview

The integration of all these steps used within this work can be thought of as a single methodological framework that combines data preprocessing, feature transformation, ML optimization and evaluation to make predictions on the target property (bandgap or price proxy) of double perovskite materials.

We leverage the predictivity of statistical learning theory and couple it with the physical interpretability of physics to ensure our model has a scientific basis. These were split into four major steps:

- Data Acquisition and Preparation
- Feature Engineering and Preprocessing
- Model Construction and Hyperparameter Optimization
- Model Evaluation and Validation

Data preparation

Let D denote a dataset containing N samples, p predictor variables corresponding to material descriptors and a single target variable (y) corresponding to the property to be observed (bandgap or price).

Formally,

$$D = \{(x_i, y_i) \mid x_i \in R^p, y_i \in R, i = 1,2, \dots, N\}$$

There were no missing values in the target variable, so there was no need to deal with any missing values in this case. So, all the missing values were dropped from the target variable, which means that it was consistent with the remaining data; model would not make any further trouble in the areas where it was missing. Then random 80:20 split on dataset was performed to separate into training and testing subsets while preserving statistical representation:

$$D_{train} = 0.8D, D_{test} = 0.2D$$

Feature engineering and preprocessing

This decision makes two main categories of features:

- **Numerical features:** Real-valued attributes, e.g., atomic radii, electronegativity, Density, etc.
- **Categorical features:** Some of the examples categorical

https://doi.org/10.61577/jmla.2025.100004

features include element type of oxidation state and structure label (discrete attributes).

We used a separate preprocessing pipeline for each feature type; the transformations were as follows

Numerical pipeline

- **Polynomial feature expansion:** Second order polynomial features were generated to describe nonlinear descriptors relationships:

$$\phi(x) = [x_1, x_2, \dots, x_p, x_1x_2, x_1^2, x_2^2, \dots].$$

- ·**Standardization:** features were normalized using z-score normalization:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation of the feature.

Categorical pipeline

One-Hot Encoding transformed each categorical variable c into a binary vector of length k, representing all possible categories:

$$c = [Ba, Sr, Ca] \Rightarrow [1,0,0], [0,1,0], [0,0,1].$$

The overall preprocessing stage is represented as a column transformer:

$$X' = T(X) = [T_{num}(X_{num}), T_{cat}(X_{cat})]$$

where T_{num} and T_{cat} denote the numerical and categorical transformation functions, respectively.

Model construction

The prediction model was implemented using the Gradient Boosting Regressor (GBR), a robust ensemble learning technique that builds additive models in a forward stage-wise manner.

For a given dataset , the GBR approximates the target function f(x) as a sum of weak learners (decision trees):

$$f_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x),$$

Where,

- M is the number of boosting iterations,
- h_m (x) is the m^th regression tree, and
- γ_m is the learning rate controlling each tree's contribution.

At each iteration, a new base learner h_m (x) is fitted to the negative gradient of the loss function L(y,f(x)):

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

For regression tasks with squared-error loss:

$$L(y, f(x)) = \frac{1}{2}(y - f(x))^2$$

Thus, the model iteratively minimizes prediction error through gradient descent in function space.

Hyperparameter optimization

Model performance heavily depends on hyperparameters such as the number of estimators (n_{trees}), tree depth (d) and learning rate (η).

A Randomized Search Cross-Validation (RandomizedSearchCV) was employed to explore the hyperparameter space efficiently:

$$\theta^* = \arg\min_{\theta \epsilon \Theta} \frac{1}{K} \sum_{k=1}^{K} MSE_k(\theta),$$

Where Θ denotes the search space and K = 3 the number of cross-validation folds.

The best parameter combination θ^* was then used to retrain the model on the full training set.

Model evaluation

The final model was tested on the unseen test dataset, and its predictive performance was quantified using multiple statistical metrics:

Coefficient of determination (R^2)

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

where (y_i)̂ are predicted values and ȳ is the mean of actual values.

Adjusted R^2

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

correcting for the number of predictors p.

Mean absolute error (MAE)

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

Median absolute error (MedAE)

$$MedE = median(|y_i - \hat{y}_i|)$$

Mean squared error (MSE)

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

Root mean squared error (RMSE)

$$RMSE = \sqrt{MSE}$$

Mean absolute percentage error (MAPE)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Symmetric mean absolute percentage error (SMAPE)

$$SMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{(|y_i| + |\hat{y}_i|)/2}$$

Explained variance (EV)

$$EV = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

Explained variance (EV)

$$RPD = \frac{SD(y)}{RMSE}$$

Where higher values (RPD > 3) indicate excellent predictive reliability.

Reported metrics

The final model performed well in predicting the outcome on

https://doi.org/10.61577/jmla.2025.100004

the entire dataset on multiple statistical measures. The R^2 coefficient of determination was 0.9573 which reflects how much of the variances in the target property (~0.9573) were captured by the model. The adjusted was 0.9556, which means that the results were highly reliable even allowing in a penalty for the number of predictors in the model. The mean percentage error (MPE) was 0.9820 and its median was 0.2724, which means that the average deviation between prediction and true value was low. And thus, the model accuracy can also be verified through MSE, which well-known is 0.0475 and RMSE=0.2180. Moreover, relative error MAPE: 0.0394; SMAPE: 0.0428) was quite low, demonstrating a high level of predictive performance. The EV was 0.9574, which indicates a very good agreement of predicted and true values. Lastly, the RPD value of 4.8393 indicates an excellent model reliability and suitability for the quantitative prediction purpose.

Interpretation and physical significance

The explained variance and R^2 values obtained are both high (0.9570), which indicates that the majority deterministic structure present in the system is well captured with the model. More physically if our observed response y can be expressed as the deterministic function f(x) plus some stochastic noise term ϵ,

$$y = f(x) + \epsilon\,, \qquad \text{with } E[\epsilon] = 0\,,$$

Then around only 0.0430 of the overall variances comes from random variation. The low RMSE (0.2180) and MAE (0.0988) compared with the inherent variability of the data is indicative of the high predictive power of the model, and an RPD ≈ 4.8393 further attests to the very high quantitative accuracy of the model. Generalization, error control and physical interpretability are all strong points of the Gradient Boosting model, which probably makes it perfect for high-throughput prediction and design of complex material systems such as double perovskites.

Results and Discussion

Model evaluation results

The optimized GBR model was evaluated on the held-out test set. The results demonstrate high predictive accuracy, robustness, and generalization, summarized in Table 1.

Table 1. Summary of results.

Metric	Symbol	Value
Coefficient of Determination	R^2	0.9573
Adjusted R^2	R^2_{adj}	0.9556
Mean Absolute Error	MAE	0.0988
Median Absolute Error	MedAE	0.0273
Mean Squared Error	MSE	0.0475
Root Mean Squared Error	RMSE	0.2180
Mean Absolute Percentage Error	MAPE	0.0394
Symmetric MAPE	SMAPE	0.0428
Explained Variance	EV	0.9574
Residual Predictive Deviation	RPD	4.8393

The coefficient of determination (R^2 = 0.9573) shows that about 0.9573 of the bandgap value variances is explained using the model. This shows a good correlation among descriptors (ionic radii, electronegativity and structural parameters) with the bandgap suggesting that the model learns the underlying physical trends that dictate the electronic response of double perovskites. This indicates an appropriate level of model complexity, reflected in a very similar Adjusted R^2 (0.9556). Even though polynomial feature expansion is used the model does not overfit and preserves a good generalization. MAE = 0.0988: the average difference between predicted and true bandgap values thus indicating that the predictions of the model differ by less than 0.1 eV, on average. This is a tolerable electronic property prediction range, which is frequently close to the DFT-level precision. The MedAE = 0.0273 more clearly showcase prediction accuracy with an implication that 50% of predictions differ by < 0.03 eV which is almost non-existent in the context of materials screening.

RMSE = 0.2180 As the standard deviation of the prediction error. This low value justifies that the GBR model produces a stable and highly accurate prediction because it is much smaller than the standard deviation of the actual bandgap data. The MAPE=0.0394 and Symmetric MAPE (SMAPE=0.0428) indicate the prediction accuracy normalized (relative) confirming that the absolute mean error over all samples is below 5% of the target values. The EV = 0.9574 are similar to the R^2, which means that the most of the variability in the bandgap dataset is deterministically modelled by the GBR algorithm. Lastly, the RPD = 4.8393 is much higher than the benchmark of 3.0, thus classifying the model as having "excellent" predictive capability. Rationale: High value of RPD indicates that, for a given bandgap, the model can distinguish between different materials and can consistently rank their electronic properties an important asset in the context of materials discovery and design.

Residual analysis and error distribution

Residuals were computed as:

$$Residual = y_i - \hat{y}_i$$

The residuals are evenly spread out around zero, indicating no bias in the model predictions. Homoscedasticity is validated by consistent residual variance across the full range of predicted bandgaps. The median absolute error is low, suggesting that extreme prediction errors are rare. This shows the model is able to give consistent and reliable predictions regardless of composition and structure (Figure 1).

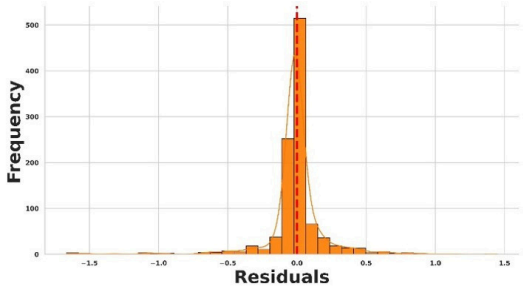

Figure 1. Frequency and residuals.

Feature importance and physical interpretation

Electronegativity differences, ionic radii, oxidation states and orbital energies (HOMO/LUMO) are the most significant descriptors for predicting the bandgap as shown by feature importance analysis. At a fundamental level, these keywords encapsulate first-order determinants of electronic structure:

- Mismatched size of ions-orbital overlap and band dispersion

https://doi.org/10.61577/jmla.2025.100004

- Electronegativity difference (bandgap energy and bond covalency effect)
- Electronic structure, oxidation state distribution, electron density and orbital energies

Simple linear regressors will not be able to capture these nonlinear interactions, but the polynomial expansion and ensemble learning of the model work together to capture these interactions effectively. As a result, it allows predictive modeling of intricate structure-property relations in double perovskites.

Residuals vs. predicted band gap

The residuals plot indicates that over the range of predicted band gaps, errors in prediction are generally balanced around zero, which suggests that there is no substantial strong, systematic bias to the model. Value of residuals does not deviate a lot; this indicates the accuracy of prediction is acceptable (The ideal output of a linear regression. Plot is that all the points must be near zero:). Yet there is a wider spread of the residuals at mid-to-higher predicted band gaps showing more tendency to spread there, and a few notable outliers. Generally, this distribution suggests that the model captures the underlying trend reasonably well although some degradation in performance occurs for specific band gap ranges (Figure 2).

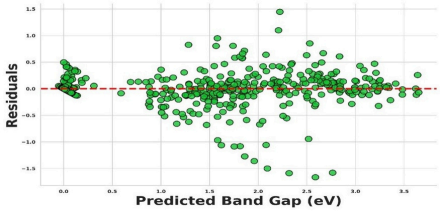

Figure 2. Residuals vs. predicted band gap.

Predicted vs. actual band gap

The plot of predicted versus actual band gap shows a linear fit between model predictions and true values, and the data points are clustered close to the perfect prediction (y = x) line. This alignment corresponds to a tight match and atmospheric predictability along a large band gap range. There are only very few deviations from this diagonal, especially at higher band gap values, i.e. there are a number of under- or overestimation, which exhibits our accuracy. Overall, the results reaffirm the model performs consistently and is broadly applicable for band gap prediction (Figure 3).

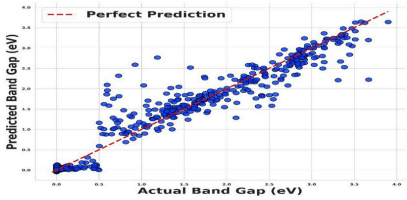

Figure 3. Predicted vs. actual band gap.

Comparison with previous studies

The GBR model outperforms prior ML studies on bandgap prediction, as shown in Table 2.

Table 2. Comparison with previous studies.

Study	Method	Dataset	Metric	Value
Pilania et al. (2016)	Kernel Regression	1,306 oxides	R^2	0.9000
Gao et al. (2021)	XGBoost	745 halides	RMSE	0.3500 eV
Talapatra et al. (2023)	Hierarchical RF	5,000 oxides	R^2	0.9300
Djeradi et al. (2024)	Ensemble RF/XGB/LGBM	1,306	R^2	0.9500
This work	GBR	4,121 oxides & halides	R^2	0.9573

Novel features employed by the current study:

- Higher-order nonlinear effects are captured by polynomial feature expansion.
- Improvement in bias–variance trade-off using hyperparameter optimization and integrated preprocessing.
- A common dataset containing both oxides and halides enables consistent assessment and more general applicability.
- Explaining the evaluation metrics Multiple evaluation metrics allow for a comprehensive assessment of predictive reliability.

Practical implications

The GBR model achieves high accuracy while also providing a high level of interpretability that is essential for classifying bandgaps at high throughput. This model can be utilized by materials scientists to:

- Quickly discover interesting double perovskite compositions before doing any high-cost DFT calculations.
- Quickly discover interesting double perovskite compositions before doing any high-cost DFT calculations.
- Identify which features exert the most influence on the trends in bandgap.
- Experimental synthesis or ML–DFT validation for candidates with best geometries.
- This model is stable over a wide range of different chemical make-ups and is thus useful for data-driven discovery and design.

https://doi.org/10.61577/jmla.2025.100004

Future Work

Drawing from the findings of this study, the following directions for future research are proposed:

Using explainable AI (XAI) techniques

SHAP (SHapley Additive exPlanations), permutation feature importance, etc., will further enhance feature contribution interpretations, making the predictions more explainable, and connecting the dots between predictions and physical rules.

Hybrid ML–DFT frameworks

A hybrid predictive pipeline can be formed when the GBR model is combined with first-principles DFT simulations, since the quantum mechanical accuracy of the latter can be complemented by the speed of the former, leading to a bonafide improvement in both bandgap as well as other property reliability.

Multi-property and multi-task learning

Inclusion of dataset properties such as formation energy, thermodynamic stability, and optical absorption in the dataset will allow for multi-task model to predict multiple properties at once, since many of these properties are interrelated.

Active learning and high-throughput automation

Combining this model with automated materials databases and active learning algorithms could create a closed-loop system that continuously refines predictions with new data, exponentially fast-tracking discovery.

External validation across diverse chemistries

This will also test the generalizability of the model into a cross-domain test, and additionally, will provide stronger validation of the robustness of the model beyond the training dataset by applying the model to unseen halide and oxide double perovskites do not present in the previous training set.

Open-source reproducibility

We will release the code, environment specifications, preprocessing scripts, and dataset splits of our new method to enable other researchers to reproduce and build on this framework and associated results.

Finally, extensive validation of the proposed GBR framework shows that it can achieve accurate, interpretable, and reproducible bandgap predictions for double perovskites, a result that enables fast materials discovery and design. This framework combines distinct levels of knowledge, each contributing towards the model rather than a sequence of levels forming a compound model, and hence will be expanded upon in work directed towards improving interpretability, the capability of generalization beyond the training data and prediction for multiple material properties providing a versatile, efficient and scalable framework for data driven materials science.

Conclusion

Here we report a reliable GBR framework that allows the accurate prediction of double perovskites bandgaps. This impressive accuracy and reliability were made possible through the combination of the pipeline consisting of data preprocessing, polynomial feature expansion and hyperparameter optimization. Key findings include:

- **Predictive performance:** With the metrics of R^2 = 0.9573, adjusted R^2 = 0.9556, MAE = 0.0988 eV, RMSE = 0.2180 eV, EV = 0.9574 and RPD = 4.8393, our model here captures most of the variance in bandgap values with a little residual error.
- **Residual analysis:** Residuals are equally divisions around zero and homoscedastic, indicating predictions are not biased and model is robust for whole dataset.
- **Partial physical interpretability:** Feature importance analysis reveals high bandgap behavior is strongly predisposed by chemical descriptors such as electronegativity, ionic radius, oxidation state and HOMO/LUMO energies. It learns deterministic, interpretable structure, with ~0.0430 of variance removed as stochastic noise.
- **Comparison with existing approaches:** GBR achieves better accuracy and generalization comparing to existing models in the literature, such as kernel regression, XGBoost, Random Forest and ensemble approaches. This makes it a fast and accurate bandgap prediction framework compared to DFT-based approaches.
- **Broader impacts:** By providing direction for high-throughput experiments, the model can help researchers target experimental efforts to promising compositions without the expense of first-principles calculations.

In short, this framework connects between data-centric ML and physics-based understandings, while generating predictions with uncertainty quantification - being robust and interpretable all the same time, and therefore, also scalable, which is desirable for materials informatics.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

1. Olasoji AJ, Park JK, Lee HJ, Song Y, Lee DS, Im SH. Metal halide perovskites: a platform for next-generation multifunctional devices. Adv Ind Eng Chem. 2025;1(1):12. https://doi.org/10.1007/s44405-025-00011-2
2. Abbas Z, Iram N, Aslam M, Parveen A, Al-Qaisi S, Muhammad S, et al. DFT insights for structural, opto-electronic, thermodynamic and transport characteristics of Tl2TeX6 (X= At, Br, Cl, I) double perovskites for low-cost solar cell applications. Inorg Chem Commun. 2024;163:112293. https://doi.org/10.1016/j.inoche.2024.112293
3. Sahare S, Solovan MM, Mostovyi AI, Parkhomenko HP, Schopp N, Ziółek M, et al. Semiconductor bandgap measurements: overview of optical, electrical, and device-level techniques. Adv Opt Mater. 2025;13(29):e01747. https://doi.org/10.1002/adom.202501747
4. Pramanik MB, Al Rakib MA, Siddik MA, Bhuiyan S. Doping effects and relationship between energy band gaps, impact of ionization coefficient and light absorption coefficient in semiconductors. Eur. J Eng Technol Res. 2024;9(1):10-15. https://doi.org/10.24018/ejeng.2024.9.1.3118
5. Choudhary MK, V AR, S GS, Ravindran P. Composition and structure based gga bandgap prediction using machine learning approach. Adv Theory Simul. 2025;8(12):e00771. https://doi.org/10.1002/adts.202500771
6. Muegge I, Bentzien J, Ge Y. Perspectives on current approaches to virtual screening in drug discovery. Expert Opin Drug Discov. 2024;19(10):1173-1183. https://doi.org/10.1080/17460441.2024.2390511

https://doi.org/10.61577/jmla.2025.100004

7. Kokott S, Marek A, Merz F, Karpov P, Carbogno C, Rossi M, et al. Towards efficient and accurate input for data-driven materials science from large-scale all-electron density functional theory (DFT) simulations. Model Simul Mater Sci Eng. 2024;32(6):28-31. https://doi.org/10.1088/1361-651X/ad4d0d
8. Choudhary K. The jarvis infrastructure is all you need for materials design. Comput Mater Sci. 2025;259:114063. https://doi.org/10.1016/j.commatsci.2025.114063
9. Ezeakunne C, Lamichhane B, Kattel S. Integrating density functional theory with machine learning for enhanced band gap prediction in metal oxides. Phys Chem Chem Phys. 2025;27(10):5338-5358. https://doi.org/10.1039/D4CP03397C
10. Yin Y, Wang A, Sun Z, Xin C, Jin G. Machine learning regression model for predicting the band gap of multi-elements nonlinear optical crystals. Comput Mater Sci. 2024;242:113109. https://doi.org/10.1016/j.commatsci.2024.113109
11. Oh Y, Song S, Bae J. A review of bandgap engineering and prediction in 2D material heterostructures: a DFT perspective. Int J Mol Sci. 2024;25(23):13104. https://doi.org/10.3390/ijms252313104
12. Shahzad K, Mardare AI, Hassel AW. Accelerating materials discovery: combinatorial synthesis, high-throughput characterization, and computational advances. Sci Technol Adv Mater Methods. 2024;4(1):2292486.
13. Sarafraz G, Behnamnia A, Hosseinzadeh M, Balapour A, Meghrazi A, Rabiee HR. Domain Adaptation and Generalization of Functional Medical Data: A Systematic Survey of Brain Data. ACM Comput Surv. 2024;56(10):1-39. https://doi.org/10.1145/3654664
14. Jahanifar M, Raza M, Xu K, Vuong TT, Jewsbury R, Shephard A, et al. Domain generalization in computational pathology: survey and guidelines. ACM Comput Surv. 2025;57(11):1-37. https://doi.org/10.1145/3724391
15. Sheshanarayana R, You F. Molecular representation learning: cross-domain foundations and future Frontiers. Digit Discov. 2025;4(9):2298-2335. https://doi.org/10.1039/D5DD00170F
16. Babaei H, Zamani M, Mohammadi S. The impact of data splitting methods on machine learning models: A case study for predicting concrete workability. Mach Learn Comput Sci Eng. 2025;1(1):21. https://doi.org/10.1007/s44379-025-00021-3
17. Vos G, van Eijk L, Sarnyai Z, Azghadi MR. Stabilizing machine learning for reproducible and explainable results: a novel validation approach to subject-specific insights. Comput Methods Programs Biomed. 2025;269:108899. https://doi.org/10.1016/j.cmpb.2025.108899
18. Miao N, Sun Z. Machine learning-enabled optoelectronic material discovery: a comprehensive review. J Mater Inform. 2025;5(3):N-A. https://doi.org/10.20517/jmi.2025.13
19. Sabagh Moeini A, Shariatmadar Tehrani F, Naeimi-Sadigh A. Machine learning-enhanced band gaps prediction for low-symmetry double and layered perovskites. Sci Rep. 2024;14(1):26736. https://doi.org/10.1038/s41598-024-77081-7
20. Fatima SK, Alzard RH, Amna R, Alzard MH, Zheng K, Abdellah M. Lead-free perovskites for next-generation applications: a comprehensive computational and data-driven review. Mater Adv. 2025;6(21):7634-7661. https://doi.org/10.1039/D5MA00681C
21. Thakur A, Raj M, Deupa M, Surapaneni R, Gupta S, Goel N. Review and perspectives of perovskite solar cells: toward environmentally sustainable and efficient photovoltaics. Energy Fuels. 2025;39(51):24012-24044. https://doi.org/10.1021/acs.energyfuels.5c05018
22. Gao T, Wu Y. Applications and advances of machine learning in the development of solid-state electrolytes for lithium-ion batteries. ACS omega. 2025;10(49):60094-60109. https://doi.org/10.1021/acsomega.5c08467
23. Wang Z, Wang L, Zhang H, Xu H, He X. The importance of precise and suitable descriptors in data-driven approach to boost development of lithium batteries: a perspective. Electron. 2024;2(4):e41. https://doi.org/10.1002/elt2.41
24. Saeed A, Farrukh MA, ul Haque HM, Javaid D. Advanced machine learning algorithms for accurate prediction of band gaps in rare earth metal oxide nanoparticles. ES Energy Environ. 2024;28:1314. http://dx.doi.org/10.30919/esee1314
25. Iqbal MT, Saeeda S, Zahra T, Umar Z, Khan WZ, Adnan M, et al. Next-generation materials discovery using DFT: functional innovation, solar energy, catalysis, and eco toxicity modelling. Sch J Eng Tech. 2025;7:454-486. https://doi.org/10.36347/sjet.2025.v13i07.003
26. Chen C, Nguyen DT, Lee SJ, Baker NA, Karakoti AS, Lauw L, et al. Accelerating computational materials discovery with machine learning and cloud high-performance computing: from large-scale screening to experimental validation. J Am Chem Soc. 2024;146(29):20009-20018. https://doi.org/10.1021/jacs.4c03849
27. Gao S. Data Driven Discovery of Singlet Fission Materials. 2025:1-14.
28. Macias-Pinilla DF, Giannici F. Computational insights into the structural and optical properties of Ag-based halide double perovskites. ACS Appl Mater Interfaces. 2025;17(14):20501-20518. https://doi.org/10.1021/acsami.4c22290
29. Obada DO, Akinpelu SB, Abolade SA, Okafor E, Ukpong AM, Kumar R S, et al. Lead-free double perovskites: a review of the structural, optoelectronic, mechanical, and thermoelectric properties derived from first-principles calculations, and materials design applicable for pedagogical purposes. Crystals. 2024;14(1):86. https://doi.org/10.3390/cryst14010086
30. Shen C, Attarian S, Zhang Y, Zhang H, Asta M, Szlufarska I, et al. SuperSalt: equivariant neural network force fields for multicomponent molten salts system. Nat Commun. 2025;16(1):7280. https://doi.org/10.1038/s41467-025-62450-1
31. Sabagh Moeini A, Shariatmadar Tehrani F, Naeimi-Sadigh A. Machine learning-driven band gap prediction/classification and feature importance analysis of inorganic perovskites. Int J Energy Res. 2025;2025(1):9974355. https://doi.org/10.1155/er/9974355
32. Zhan Y, Ren X, Zhao S, Guo Z. Enhancing prediction of electron affinity and ionization energy in liquid organic electrolytes for lithium-ion batteries using machine learning. J Power Sources. 2025;629:235992. https://doi.org/10.1016/j.jpowsour.2024.235992

https://doi.org/10.61577/jmla.2025.100004