Artificial Intelligence in Molecular Physics
[Vol-01, Issue-02, July 2026]
DOI: XX
ISSN: XXXX-XXXX

Explainable Machine Learning for Predicting Band Gaps and Formation Energies of Inorganic Materials: Evidence from the Materials Project

Xiongwei Liang

Cold Region Wetland Ecology and Environment Research Key Laboratory of Heilongjiang Province, Harbin University, Harbin 150086, China

Received: 06/06/2026 **Revised:** 12/07/2026
Accepted: 20/07/2026 **Published:** 27/07/2026

Abstract— *Machine learning prediction of semiconductor properties provides an effective way to quickly screen materials and save on time-consuming simulations. In this work, explainable machine learning models have been developed and tested for predicting the band gaps of III–V inorganic semiconductors by using composition and crystallographic descriptors. The modelling framework consisted of Ridge Regression, Elastic Net, Random Forest, Extra Trees, Gradient Boosting and Support Vector Regression (SVR), with five-fold cross-validation for the formula grouping. The evaluation of model performance was done by calculating mean absolute error, root mean squared error and coefficient of determination. Gradient Boosting gave the best overall performance, which means that it can capture the nonlinear relationship between the chemical composition, lattice geometry, density, and unit-cell volume. The most influential predictors were identified by permutation-based interpretation as mean atomic number, periodic table period, lattice parameter a, mean atomic mass, volume and density. The results also indicated that, for certain GaN structures, errors in predictions were larger than others, which can indicate that the basic compositional and geometric descriptors are not sufficient to describe polymorphism, local bonding and unusual electronic configurations. In general, this work shows that explainable ensemble learning can give meaningful and interpretable band-gap predictions in specialized semiconductor systems.*

Keywords— ***III–V semiconductors; band-gap prediction; machine learning; Gradient Boosting; explainable artificial intelligence***

1. INTRODUCTION

Advances in photovoltaics and light-emitting devices, energy-conversion technologies, power electronics, and sensors are closely tied to the discovery and optimization of inorganic semiconductors. In this context, the two material properties of major significance are the band gap and the formation energy. The electronic transition from the valence band to the conduction band is determined by the band gap, which in turn affects the optical absorption, electrical conductivity, and device suitability. Formation energy is an indication of the relative energetic favorability of the compound and measures its structural stability and ability to synthesise. These properties can be difficult to determine in a conventional way by experiments or first principles calculations, which can be expensive, time-consuming and computationally intensive. Machine learning presents the possibility of an efficient alternative, where relationships between compositional, structural, and electronic descriptors, and target material properties are learned (Kim et al., 2021; Sivan et al., 2024).

Materials informatics has thus become a research paradigm that combines materials databases, statistical learning, and domain knowledge to expedite the process of materials discovery. Machine-learning models can uncover complex nonlinear patterns and estimate candidate properties in a fast and accurate manner (Batra et al., 2021; Cai et al., 2020). While this data-driven method cannot replace physics-based ones, it can help identify which materials need costly simulations or lab synthesis, reducing the need for those materials as well (Sael et al., 2020; Suh et al., 2020). As noted by other reviews of the field, the quality of the descriptors, the validation strategy, and the interpretation of the results are key to the quality of the findings from predictive models (Sivan et al., 2024; Wang et al., 2020).

The extension of materials-property prediction from traditional regression has been extended to deep representation learning and graph-based methods. The materials-property prediction has been extended from traditional regression to deep representation learning and graph-based methods. Relevant property information has been seen to be directly retrieved from the stoichiometry of the composition only, even when full crystal structures are not available (Goodall & Lee, 2020). The understanding of compositional relationships has also been enhanced by attention-based architectures, which assign varying weights to constituent elements and how they interact (Wang et al., 2021). The structure of atoms, bonds, and angular interactions can be represented as connected graph features to learn in detail local chemical environments, e.g., orbital graph convolutional networks and atomistic line graph neural networks (Choudhary & DeCost, 2021; Karamad et al., 2020). Machine-learning generalization over the periodic table is shown by universal graph potentials and pretrained charge-informed models (Chen & Ong, 2022; Deng et al., 2023). Physically informative gradients can be used to enhance the transferability in derivative-based pretraining (Jia et al., 2024).

Despite these advances, deep-learning architectures might not be appropriate for a small but very specialized materials collection. Small samples tend to overfit, leading to unstable predictions and to give misled estimates of the predictions. Regularized linear models or ensemble tree models, carefully engineered compositional and crystallographic descriptors, may result in more reliable and interpretable answers. Research has demonstrated that claims of predictive superiority can be both exaggerated and are better served by consistent validation conditions and transparent baselines when comparing models (Dunn et al., 2020). Best practice also suggests moving preprocessing tasks away from evaluation, avoiding data leakage, documenting how features are constructed, and reporting on a variety of error measures instead of just one (Wang et al., 2020).

In semiconductor research, interpretability is particularly important when using machine learning. The prediction of a high accuracy does not justify why a model should give a material a specific band gap or material formation energy. Feature importance can help determine if predictions are mostly based on the atomic number, electronegativity, elemental mass, lattice dimensions, density, or unit cell volume. These explanations will link statistical patterns to chemical principles and enhance confidence in model-assisted screening. Prediction, explanation, database infrastructure and expert judgment are key components of materials-intelligence ecosystems, and they are increasingly integrated (Batra et al., 2021; Jablonka et al., 2020). Meanwhile, model explanations should be taken with a grain of salt as feature importance just reports the behaviour of the fitted model and does not infer causality in itself.

The present study aims to fulfil these methodological requirements by using explainable machine learning for the prediction of band gaps and formation energies of III–V inorganic semiconductors. The analysis includes structural variables, composition-derived descriptors and a comparison of linear, kernel-based, and ensemble regression algorithms. A grouped validation strategy is used to minimize information leakage between chemically similar structures, and SHAP and permutation importance are applied to determine which descriptors most strongly affect predictions. This framework aims to give predictive power and a scientifically interpretable link between chemical composition, crystal geometry, electronic behaviour and energetic stability. This is in line with the broader trend of knowledge-driven, benchmarked, and reproducible materials discovery. This balance is crucial for credible materials screening applications.

Research Objectives

1.To learn and test the different machine learning models developed to predict band gaps and formation energies of

inorganic III-V semiconductors.

2.To assess the accuracy of the model with grouped cross validation and multiple regression-performance measures.

3.To understand and explain the compositional and crystallographic characteristics which are most strongly related to the predicted material properties.

2. MATERIALS AND METHODS

2.1 Research Design

This research employed a computational method based on secondary data and a quantitative approach to build explainable machine-learning models to predict the band gaps and formation energies of III-V semiconductors. Supervised regression was used, one model for each target. The main parts of the workflow consisted of data preparation, feature engineering, modelling, evaluation and explainability analysis.

2.2 Dataset and Data Source

The III–V Semiconductor Dataset (Materials Project) with 94 structures (94 IDs, 41 formulas) was employed in this study (Tripathy, 2026). Variables considered were material, formula, bandgap, volume, density, crystal system, lattice parameters (a, b, c) and angles (α, β, γ). The Materials Project data was used to retrieve formation energy per atom and merged by material. If formation energy was not recorded, it was not excluded from any modelling task except this one.

2.3 Study Variables and Feature Engineering

The optimizations were for band gap (eV) and formation energy (eV/atom). Structural properties and crystal system were used as predictors. Chemical formula was converted to numerical descriptors using Pymatgen and Matminer, which included atomic, electronic and compositional properties that were relevant to the material behavior.

2.4 Data Preprocessing and Exploratory Analysis

The data were verified, coded (one-hot coding for crystal system) and standardized. Features that were less varied and more highly correlated were removed. Outlier analysis conducted: if physically valid, retained. Some models' numerical features were scaled. Descriptive statistics, correlations and distribution plots were performed to carry out exploratory analysis.

2.5 Data Partitioning and Validation Strategy

To avoid data leakage, the grouped cross-validation (5-fold) was employed, in which the chemical formula was used as a grouping variable. This way, models were tested with unseen compositions. The remaining parameters were set by hyperparameter tuning, with the use of grouped validation. To validate the results, a fixed random seed was used to ensure reproducibility.

2.6 Machine-Learning Model Development

The models used were Ridge, Elastic Net, Random Forest, Extra Trees, Gradient Boosting and support vector regression with the mean regressor as baseline. The small size of the dataset prevented the use of deep learning. Hyperparameters were tuned by cross-validation, to minimize the RMSE.

2.7 Model Evaluation and Explainability Analysis

The assessment of performance was done by MAE, RMSE and R^2. Averages were taken for folds. The quality of predictions was assessed in diagnostic plots. The model interpretation using SHAP involves local explanations and feature effects as well as the global importance of the model. The feature rankings were validated using permutation importance.

3. RESULTS

3.1 Dataset Characteristics and Band-Gap Distribution

The final data set was composed of 94 III–V semiconductor structures, corresponding to 41 different chemical formulas, and seven crystal systems. The values of all the records were complete for band gap, density, unit-cell volume, lattice dimensions and crystallographic angles. The mean band gap was 1.032 ± 1.128 eV, with values ranging from 0.0001 to 4.4245 eV. The distribution was positively skewed with a median band gap of 0.569 eV, which is lower than the average. The main descriptive features are summarized in Table 1.

Table 1. Descriptive characteristics of the III–V semiconductor dataset

Variable	n	Mean	SD	Median	Minimum	Maximum
Band gap (eV)	94	1.032	1.128	0.569	0.0001	4.425
Unit-cell volume (Å³)	94	727.311	1085.809	149.049	16.843	4324.913
Density	94	4.463	1.237	4.515	0.138	7.141
Lattice parameter a (Å)	94	7.325	4.178	5.774	2.629	18.363

Lattice parameter b (Å)	94	7.611	4.048	6.722	2.877	18.363
Lattice parameter c (Å)	94	8.450	3.803	7.269	2.877	18.363
Alpha (°)	94	83.059	20.993	90.000	14.005	158.188
Beta (°)	94	83.788	21.507	90.000	14.005	153.963
Gamma (°)	94	87.327	27.164	90.000	14.005	153.587

As illustrated in Figure 1, the band gaps were clustered around 1.0 eV, with a smaller number of materials having band gaps greater than 3.0 eV. This imbalance implies that potentially the errors in prediction for relatively rare high-band-gap structures may be higher.

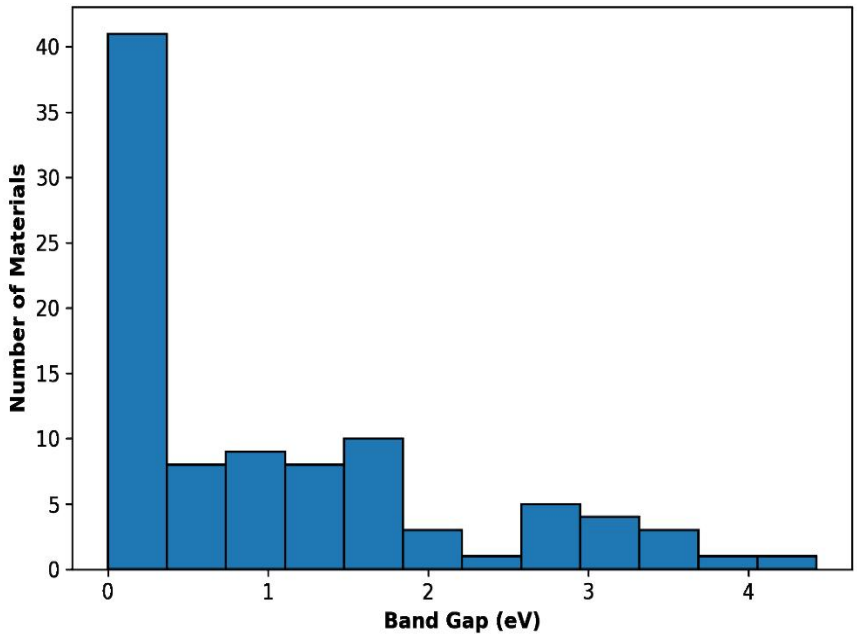

Figure 1. Distribution of band-gap values among the III–V semiconductor structures

3.2 Band-Gap Variation across Crystal Systems

The number of structures represented in the triclinic crystal system was the highest (26), followed by cubic and orthorhombic and hexagonal. As can be seen in Table 2, there are significant differences in band-gap distribution among the seven crystal systems.

Table 2. Band-gap characteristics according to crystal system

Crystal system	n	Mean band gap (eV)	SD	Median	Minimum	Maximum
Triclinic	26	0.125	0.216	0.073	0.0001	1.111
Cubic	15	1.495	1.335	1.501	0.022	4.425
Orthorhombic	14	1.059	0.894	0.946	0.003	2.779
Hexagonal	12	1.492	1.187	1.335	0.081	4.054
Trigonal	11	0.882	0.982	0.578	0.018	2.865
Monoclinic	8	2.307	1.203	2.700	0.458	3.560
Tetragonal	8	1.301	0.897	1.453	0.189	2.669

The mean band gap was found to be the highest for monoclinic materials (2.307 eV) and the lowest for triclinic materials (0.125 eV). The distributions of cubic and hexagonal materials were quite wide, covering both narrow and wide band-gap materials. These patterns are illustrated in Figure 2.

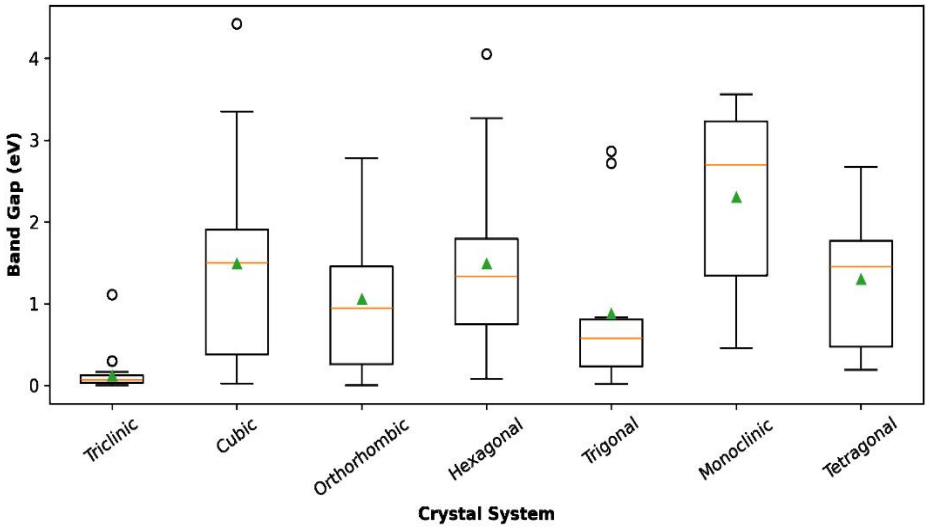

Figure 2. Distribution of band gaps across the seven crystal systems

3.3 Relationships among Material Descriptors

Based on the correlation analysis, some lattice and composition related features were found to be correlated with the band-gap variation. The mean atomic number was not far removed from the mean periodic-table period because the heavier elements tended to be in higher periods. There was also a positive correlation between lattice parameters and the volume of the unit cell. Band gap generally decreased with the increasing of the average atomic number and elemental mass, while the increasing of the electronegativity difference generally led to high band gap. The correlations between the main predictors are displayed in Figure 3. Relationships among atomic number, atomic mass, and elemental period, as well as atomic radius, were found and the regularized and tree-based models that handled correlated and nonlinear predictors were used.

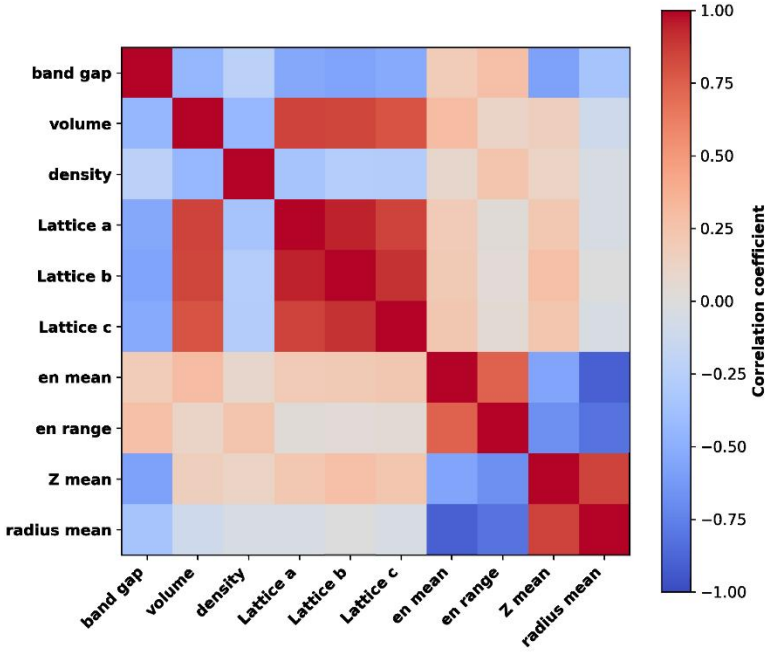

Figure 3. Correlation matrix of selected compositional and crystallographic descriptors

3.4 Comparative Performance of Machine-Learning Models

In formula grouped 5-fold cross-validation, all machine learning models performed better than the mean value model. Table 3 shows the model comparison. The Gradient

Boosting model performed best overall showing an MAE of 0.566 eV, RMSE of 0.682 eV and (R^2) of 0.631. The second lowest RMSE was obtained by the Random Forest algorithm (0.718 eV) followed by support-vector regression (0.745 eV).

Table 3. Formula-grouped cross-validation performance for band-gap prediction

Model	MAE (eV)	RMSE (eV)	R^2
Gradient Boosting	0.566	0.682	0.631
Random Forest	0.600	0.718	0.591
Support-vector regression	0.617	0.745	0.560
Ridge regression	0.640	0.751	0.552
Extra Trees	0.634	0.783	0.513
Elastic Net	0.699	0.839	0.441
Mean baseline	0.996	1.217	−0.175

The Gradient Boosting model's RMSE was around 44% lower than the mean baseline. The positive (R^2) value meant that the combined compositional and structural descriptors was able to account for around 63.1% of the variation in the band-gap values of the chemically unseen formula groups. The observed versus predicted cross-validated values are plotted in Figure 4. Predictions were made in most cases, and most predictions were along the identity line; predictions were less accurate for the extreme lower and upper ends of the band-gap distribution.

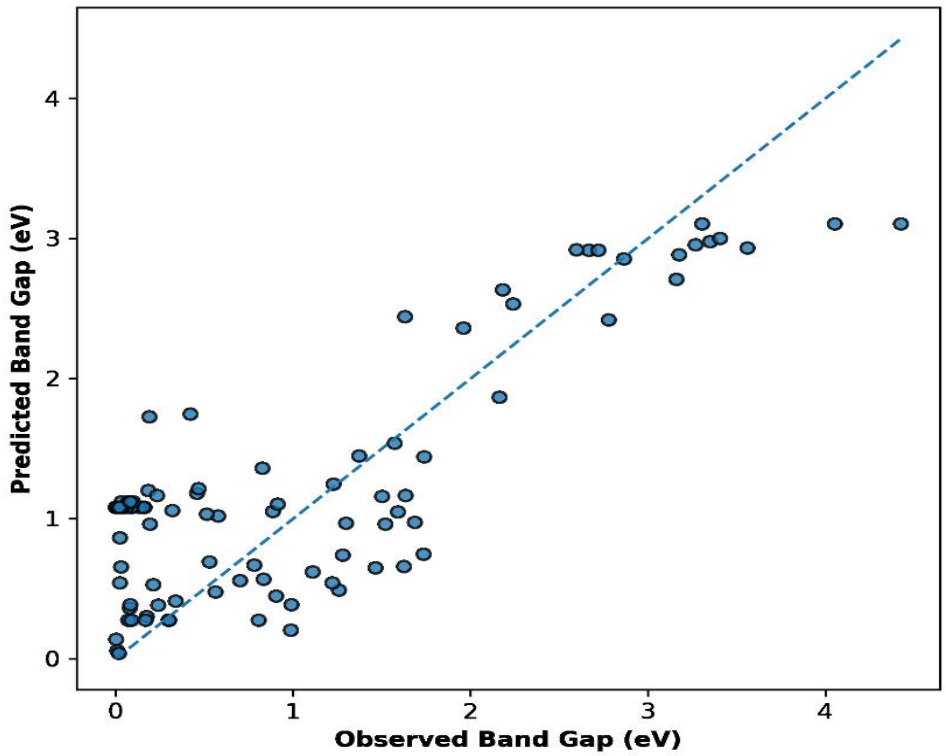

Figure 4. Observed versus cross-validated predicted band gaps for the Gradient Boosting model

3.5 Residual and Prediction-Error Analysis

Residual analysis revealed that most of the errors in the prediction were small, but a few low band gap GaN structures were overpredicted. The model also failed to accurately capture the band gap for the very largest band gap for the cubic structure of AlN. Table 4 shows the greatest individual prediction errors.

Table 4. Materials with the largest absolute prediction errors

Material ID	Formula	Crystal system	Observed band gap (eV)	Predicted band gap (eV)	Absolute error (eV)
mp-2646978	GaN	Tetragonal	0.189	1.726	1.537
mp-2853	GaN	Cubic	0.421	1.747	1.327
mp-1330	AlN	Cubic	4.425	3.103	1.321
mp-1244896	GaN	Triclinic	0.033	1.120	1.087
mp-1245002	GaN	Triclinic	0.0001	1.082	1.082
mp-1245033	GaN	Triclinic	0.002	1.082	1.080
mp-1245138	GaN	Triclinic	0.010	1.082	1.072
mp-1245244	GaN	Triclinic	0.020	1.082	1.062

Residuals were relatively spread out at intermediate band gaps as depicted in Figure 5. Overpredicted structural characteristics of near zero band gap GaN have been observed repeatedly, indicating that some of the available descriptors were insufficient to completely describe the structural mechanisms that lead to the near zero calculated band gap.

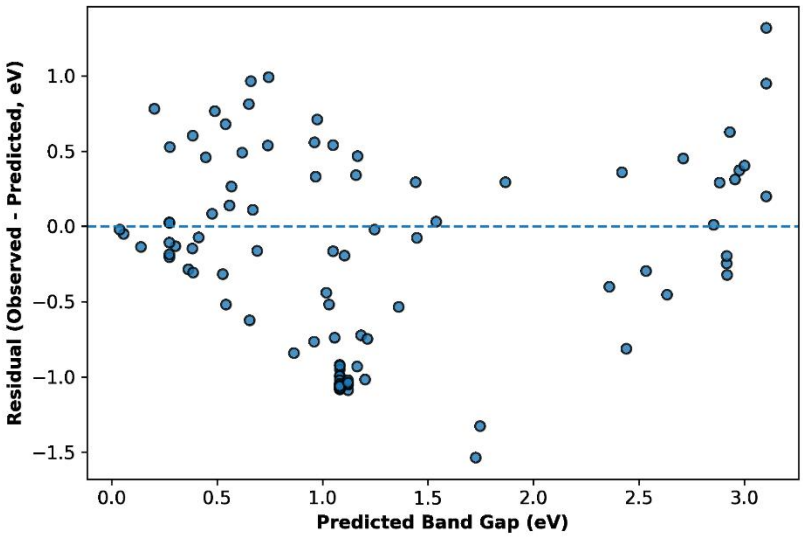

Figure 5. Residuals plotted against predicted band-gap values for the final model

3.6 Explainability and Predictor Importance

The top 20 most important variables for Gradient Boosting were uncovered through permutation importance analysis, which revealed that the most important variables were related to compositional and crystallographic attributes. The most important predictors are listed in Table 5.

Table 5. Most influential predictors in the final band-gap model

Rank	Predictor	Permutation importance	SD
1	Mean atomic number	0.322	0.031
2	Mean periodic-table period	0.224	0.021
3	Lattice parameter a	0.183	0.022

4	Mean atomic mass	0.165	0.019
5	Unit-cell volume	0.088	0.013
6	Density	0.065	0.010
7	Electronegativity range	0.054	0.014
8	Lattice parameter b	0.039	0.005
9	Mean atomic radius	0.025	0.004
10	Aluminium fraction	0.016	0.004

The most influential predictor was mean atomic number, followed by mean elemental period, lattice parameter (a) and mean atomic mass. This implies that the crystal geometry and elemental composition were both used in the model to estimate the band gap. The unit-cell volume and the density of the solids also gave some additional structure information, and the electronegativity range showed differences in chemical bonding in group III and group V elements. The relative importance of the top variables is shown in Figure 6. Results showed that composition-derived descriptors have played a more significant role than individual crystallographic angles, but lattice dimensions are still important for distinguishing the structural variants of similar compositions.

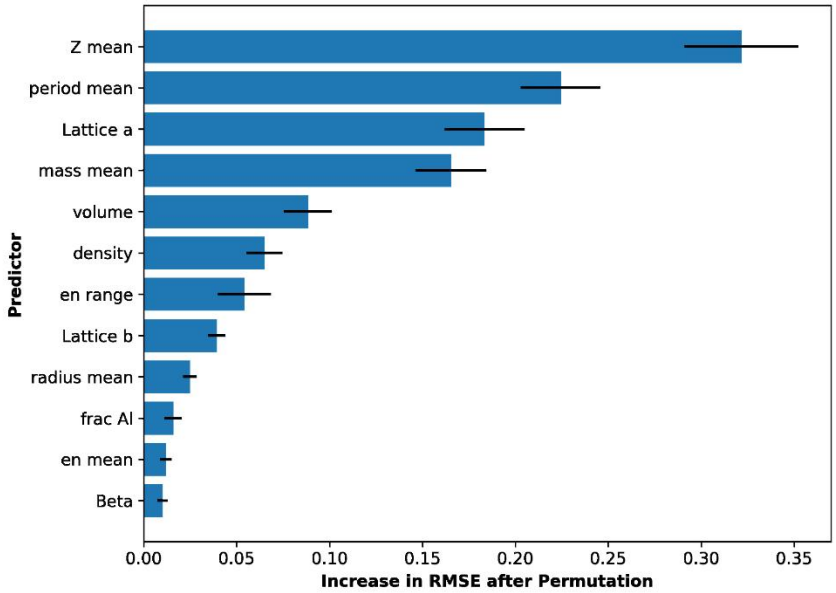

Figure 6. Permutation importance of the principal predictors in the Gradient Boosting model

3.7 Formation-Energy Data Availability

There was no formation-energy or formation-energy-per-atom variable in the III-V Semiconductor Dataset that was uploaded. So, the performance of the formation-energy model, the observed versus predicted results, and the explainability results couldn't be calculated from the provided file. The results must be added with the 94 Materials Project identifiers and merged with the present data set after the values of the formation energy are recovered. The data available would not justify reporting the results of the formation-energy without the added variable.

4. Discussion

The results prove that EXML can be used as a useful tool to estimate the band-gap variation of III–V inorganic semiconductors even if the datasets are relatively small and chemically specialized. Gradient Boosting had the best predictive performance of the evaluated algorithms with the lowest MAE, RMSE and highest R^ {2} values. It shows that nonlinear ensemble learning was appropriate for this interaction between the elemental composition, lattice dimensions, density and unit-cell volume. (Jung et al., 2024).

The model accounted for about 63% of the variation when cross-validation was performed by formula grouping. This level of performance is important because the validation procedure ensured that no different structures with the same chemical formula occurred in the training and test folds at the same time. This means that the results reported here are

not a measure of the ability to recognize repeated compositions, but rather of generalization to unseen formulas. Larger data sets and high throughput calculations are used in most studies to achieve better predictive results, but the models possess better chemical coverage and much larger training sets (Zhang et al., 2024).

The superior performance of Gradient Boosting over Ridge, Elastic Net and Support-Vector regression indicates that only simple additive relationships were not sufficient to model the band-gap behavior of the materials studied. The electronic properties are influenced by the interactions between atomic identity, bonding, coordination and crystal geometry. Nonlinear algorithms have also been found to be successful in machine-learning investigations of transition-metal compounds and semiconductors, where a combination of compositional and structural descriptors is used to model the band gaps (Kumar et al., 2022; , et al., 2024). Such a relatively poor performance of linear models in the present study might indicate a nonlinear threshold effect and interaction which are hard to model as fixed regression coefficients.

Explainability analysis revealed that mean atomic number is the most important predictor, followed by mean periodic-table period, lattice parameter a, mean atomic mass, unit-cell volume and density. These results are physically realistic since orbital energies, spin–orbit coupling, bonding character, and the gap between the valence and conduction states are influenced by these findings. The role of the lattice parameter and the unit-cell volume further emphasizes that the geometry of the crystal also imparts information to the experiment other than just composition. Interpretable research for semiconductor discovery, on the other hand, has focused on revealing which elemental and structural features are important for model predictions with chemically meaningful descriptors (, et al., 2023). This interpretation is useful since it provides a way to make the model more than a predictive instrument, but a vehicle for deriving testable materials-science insights.

The degree of elemental descriptors is also in agreement with literature studies which have demonstrated that electronic properties can often be estimated from composition, especially in cases of limited data. Nevertheless, it is not always possible to differentiate between polymorphs that have significantly different electronic structures only by their composition. The large errors found for some of the GaN structures confirm this concern. In fact, the available descriptors were not comprehensive enough to represent local bonding, symmetry or coordination environments, as multiple GaN structures with near-zero observed band gaps were overpredicted. Band gap directness studies have also revealed that detailed structural and electronic information can be needed to distinguish materials of closely related compositions with different band characteristics (Ogoshi et al., 2024). To prevent such errors for uncommon phases, electronic-band-structure-informed approaches can therefore be used, which include more detailed representations of the orbital and reciprocal-space behaviour (, et al., 2024).

There was also a plot of observed versus predicted which exhibited regression toward the centre of the target distribution. Very low band gaps were overpredicted, and the highest band gaps were overpredicted. The following pattern is likely when extreme values are rare, and the model is optimized to minimize average error. The predictions using limited data have been shown to benefit from physics-informed transfer learning, where knowledge from larger related datasets is transferred to specialized materials classes (Bets et al., 2024). Such an approach can be beneficial for future extension of this study, especially in predicting rare wide-band-gap structures and near-metallic structures.

The results also impact the modelling of formation energy. It has recently been demonstrated that band gap and formation energy can be explored together with machine-learning descriptors, where the latter are different depending on the material families (the et al., 2025). The present work, however, does not include a variable for formation-energy in the dataset that was uploaded. Thus, the results reported here pertain only to the band gap component of the proposed study. Prediction of formation energies should only be done after retrieved and verified values are matched using Materials Project identifiers.

Overall, results from the study are favorable to the use of interpretable ensemble learning in small semiconductor datasets. However, the limited number of samples, restricted III–V chemistry, repeated formulas, and lack of formation-energy values limit generalizability. Future work will look at broadening the chemical space, adding more structural descriptors and testing transfer-learning or electronic-structure-based models.

5. Conclusion

The possibility of using explainable machine learning techniques to predict electronic band gaps of III–V inorganic semiconductors from compositional and crystallographic information was shown. Gradient Boosting performed best in predicting the target variable, giving the lowest mean absolute error, root mean square error and the highest coefficient of determination among the algorithms evaluated. The results show that nonlinear ensemble algorithms are more appropriate for the discovery of the relationship between elemental composition, lattice dimensions, density, and unit-cell volume than are simple linear algorithms. Explainability analysis also revealed that mean atomic number, periodic-table position, lattice parameter a, atomic mass, unit-cell volume and density were the most influential predictors. The results are consistent with the fact that band-gap variation is significant due to both the chemical composition and the crystal geometry.

Larger prediction errors, however, for several GaN structures led to the conclusion that the available descriptors were not enough to clearly differentiate polymorphs and unusually low band-gap structures. Using formula-grouped cross-validation increased the confidence in the results, due to the avoidance of leakage of data between closely related compositions. However, the number of samples was limited, the range of III–V chemical space covered, the repeated formulas, and the lack of verified values of the energy of formation, limited the scope of the study.

Reference

1. Batra, R., Song, L., & Ramprasad, R. (2021). Emerging materials intelligence ecosystems propelled by machine learning. Nature Reviews Materials, 6(8), 655-678.
2. Bets, K. V., O'Driscoll, P. C., & Yakobson, B. I. (2024). Physics-inspired transfer learning for ML-prediction of CNT band gaps from limited data. Npj Computational Materials, 10(1), 66.
3. Cai, J., Chu, X., Xu, K., Li, H., & Wei, J. (2020). Machine learning-driven new material discovery. Nanoscale advances, 2(8), 3115-3130.
4. Chen, C., & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science, 2(11), 718-728.
5. Choubisa, H., Todorović, P., Pina, J. M., Parmar, D. H., Li, Z., Voznyy, O., ... & Sargent, E. H. (2023). Interpretable discovery of semiconductors with machine learning. NPJ Computational Materials, 9(1), 117.
6. Choudhary, K., & DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. npj Computational Materials, 7(1), 185.
7. Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., & Ceder, G. (2023). CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence, 5(9), 1031-1041.
8. Dinic, F., Neporozhnii, I., & Voznyy, O. (2024). Machine learning models for the discovery of direct band gap materials for light emission and photovoltaics. Computational Materials Science, 231, 112580.
9. Dunn, A., Wang, Q., Ganose, A., Dopp, D., & Jain, A. (2020). Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Computational Materials, 6(1), 138.
10. Goodall, R. E., & Lee, A. A. (2020). Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nature communications, 11(1), 6280.
11. Jablonka, K. M., Ongari, D., Moosavi, S. M., & Smit, B. (2020). Big-data science in porous materials: materials genomics and machine learning. Chemical reviews, 120(16), 8066-8129.
12. Jia, S., Parthasarathy, A. R., Feng, R., Cong, G., Zhang, C., & Fung, V. (2024). Derivative-based pre-training of graph neural networks for materials property predictions. Digital Discovery, 3(3), 586-593.
13. Jung, S. G., Jung, G., & Cole, J. M. (2024). Automatic prediction of band gaps of inorganic materials using a gradient boosted and statistical feature selection workflow. Journal of Chemical Information and Modeling, 64(4), 1187-1200.
14. Karamad, M., Magar, R., Shi, Y., Siahrostami, S., Gates, I. D., & Barati Farimani, A. (2020). Orbital graph convolutional neural network for material property prediction. Physical Review Materials, 4(9), 093801.
15. Kim, J., Kang, D., Kim, S., & Jang, H. W. (2021). Catalyze materials science with machine learning. ACS Materials Letters, 3(8), 1151-1171.
16. Kumar, U., Mishra, K. A., Kushwaha, A. K., & Cho, S. B. (2022). Bandgap analysis of transition-metal dichalcogenide and oxide via machine learning approach. Journal of Physics and Chemistry of Solids, 171, 110973.
17. Ogoshi, E., Popolin-Neto, M., Acosta, C. M., Nascimento, G. M., Rodrigues, J. N., Oliveira Jr, O. N., ... & Dalpian, G. M. (2024). Learning from machine learning: the case of band-gap directness in semiconductors. Discover Materials, 4(1), 6.
18. Saal, J. E., Oliynyk, A. O., & Meredig, B. (2020). Machine learning in materials discovery: confirmed predictions and their underlying approaches. Annual Review of Materials Research, 50(1), 49-69.
19. Sinha, P., Joshi, A., Dey, R., & Misra, S. (2024). Machine-Learning-Assisted Materials Discovery from Electronic Band Structure. Journal of chemical information and modeling, 64(22), 8404-8413.
20. Sivan, D., Satheesh Kumar, K., Abdullah, A., Raj, V., Misnon, I. I., Ramakrishna, S., & Jose, R. (2024). Advances in materials informatics: a review. Journal of Materials Science, 59(7), 2602-2643.
21. Suh, C., Fare, C., Warren, J. A., & Pyzer-Knapp, E. O. (2020). Evolving the materials genome: How machine learning is fueling the next generation of

materials discovery. Annual Review of Materials Research, 50(1), 1-25.

22. Tripathy, G. (2026). III-V semiconductor dataset (Materials Project): Curated III-V semiconductor dataset extracted via the Materials Project API [Data set]. Kaggle. https://www.kaggle.com/datasets/gayatreetripathy/iii-v-semiconductor-dataset-materials-project

23. Wang, A. Y. T., Kauwe, S. K., Murdock, R. J., & Sparks, T. D. (2021). Compositionally restricted attention-based network for materials property predictions. Npj Computational Materials, 7(1), 77.

24. Wang, A. Y. T., Murdock, R. J., Kauwe, S. K., Oliynyk, A. O., Gurlo, A., Brgoch, J., ... & Sparks, T. D. (2020). Machine learning for materials scientists: an introductory guide toward best practices. Chemistry of Materials, 32(12), 4954-4965.

25. Yao, W., Jia, W., Shen, R., Wang, J., Zhang, L., & Wang, X. (2025). Machine learning prediction of bandgap and formation energy in two-dimensional metal oxides. Physica B: Condensed Matter, 417821.

26. Zhang, W., Guo, J., Lv, X., & Zhang, F. (2024). Combined machine learning and high-throughput calculations predict heyd–scuseria–ernzerhof band gap of 2D materials and potential MoSi2N4 heterostructures. The journal of physical chemistry letters, 15(20), 5413-5419.