Prediction of Bandgap, Formation Energy, and Bulk Modulus of IIIA–VA Group Materials Using Machine Learning (XGBoost)

Abstract

We developed a robust machine-learned model using eXtreme Gradient Boosting (XGBoost) to accurately predict key electronic and mechanical properties of IIIA-VA semiconductor materials. This study addresses the significant challenge of rapidly screening new compounds, which are foundational materials for modern LEDs, solar cells, and high-speed electronics. We explored the limitations of traditional property calculation methods, specifically Density Functional Theory (DFT), emphasizing the necessity for a more effective and computationally efficient solution to overcome the time and resource bottleneck that hinders high-throughput materials discovery. We also discussed the machine learning procedure used to construct our predictive model, which was trained on a dataset of 254 unique IIIA-VA compounds sourced from the Materials Project, using a comprehensive set of 145 elemental features. Furthermore, we address techniques for evaluating the performance and robustness of the model on unseen datasets using the coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE).. Using this trained model, we investigated its ability to predict bandgap, formation energy, and bulk modulus. On the test set, we obtained high accuracy with an R^2 of 0.949 for bandgap (MAE = 0.2689 eV, RMSE = 0.4632 eV) and 0.946 for bulk modulus (MAE = 16.971 GPa, RMSE= 26.446 GPa), along with a strong predictive R^2 of 0.736 for formation energy (MAE = 0.153 eV/atom, RMSE = 0.334 eV/atom). Our results demonstrate significant improvements in prediction speed compared to prior DFT studies, highlighting the effectiveness of XGBoost in achieving high precision with minimal computational cost. This study offers a comprehensive analysis and a validated ML workflow, delivering valuable insights into the property landscape of IIIA-VA materials and paving the way for future rapid screening and design in materials science.

Keywords: Machine Learning, XGBoost, IIIA-VA Semiconductors, Materials Informatics, Density Functional Theory (DFT), Bandgap, Formation Energy, Bulk Modulus, Materials Project, High-Throughput Screening.

Introduction

IIIA-VA semiconductor materials, compounds formed by combining elements from Group IIIA (e.g., Aluminum, Gallium, Indium) and Group VA (e.g., Nitrogen, Phosphorus, Arsenic), are foundational to modern solid-state technology. These materials, which include well-known examples such as Gallium Arsenide (GaAs), Indium Phosphide (InP), and Gallium Nitride (GaN), are covalent crystals that often crystallize in zinc-blende or wurtzite lattice structures. A key characteristic of this class is their diverse and tunable electronic structure, exhibiting either direct or indirect band gaps. This unique property makes them indispensable for a vast range of applications, including high-efficiency Light-Emitting Diodes (LEDs), laser diodes, high-speed and RF electronics, multi-junction solar cells, and infrared detectors.

The immense potential of IIIA-VA materials in these diverse applications hinges on the precise tuning of their fundamental properties. The discovery of new compounds with specific, desired characteristics—for instance, a specific bandgap for a new LED color or high bulk modulus for a durable component—is a primary goal of modern materials science [3, 5]. This project focuses on three such critical properties:

Bandgap (eV): The energy difference between the valence and conduction bands, which dictates the material's optical and electronic behavior, including the color of light it can emit or absorb.

Formation Energy (eV/atom): A measure of a compound's thermodynamic stability, which determines if it can be synthesized and will remain stable against decomposition.

Bulk Modulus (GPa): A fundamental mechanical property indicating the material's resistance to compression or "stiffness." An accurate and efficient examination of these properties is essential before a material can be considered for costly laboratory synthesis and application.

For decades, the "gold standard" for investigating the characteristics of new materials has been Density Functional Theory (DFT). DFT-based *ab initio* quantum mechanics calculations are known for their high accuracy, providing a first-principles understanding of the electronic structure and related properties of materials. This method has been so successful that it forms the backbone of large-scale open-access databases, most notably The Materials Project [1], which has computed and cataloged properties for hundreds of thousands of materials. These calculations allow researchers to explore the property landscape *in silico* (via computation) before attempting to create a material in a lab.

Despite its accuracy, the DFT method encounters significant challenges that form a major bottleneck in materials discovery. The primary drawback is the high computational cost. DFT calculations are extremely resource-intensive, requiring significant computational power (often on supercomputing clusters) and time, frequently spanning hours or even days for a single compound's property calculations. This limitation renders DFT fundamentally impractical for the rapid screening of massive structural libraries or the exploration of vast compositional spaces (e.g., ternary and quaternary IIIA-VA alloys). Given the near-infinite number of possible elemental combinations, this computational constraint fundamentally hinders the pace of materials discovery [6].

These constraints have inspired researchers to find a new approach, bridging the significant gap between high-accuracy but resource-intensive DFT simulations and the urgent need for rapid, large-scale screening. This has led to the rise of Machine Learning (ML) in materials informatics [3, 5]. Instead of calculating properties from first principles, an ML model learns the complex, non-linear relationships between a material's composition (represented as a set of features) and its DFT-calculated properties. By training on large, existing DFT-calculated datasets like those from the Materials Project, ML models can act as a "surrogate" for DFT, blending its high accuracy with computational efficiency.

This ML-based approach enables what DFT cannot: high-throughput screening [4, 6]. Once trained, an ML model can predict the properties of thousands of candidate materials in seconds, allowing researchers to quickly identify a small list of the most promising compounds for more rigorous DFT validation or experimental synthesis.

Several machine learning algorithms have been applied to this problem, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and tree-based ensemble methods. ANNs, while powerful, often require vast datasets and complex hyperparameter tuning to avoid overfitting. In this context, gradient-boosted decision trees have shown exceptional promise. The current work employs XGBoost (eXtreme Gradient Boosting) [2], a highly efficient and scalable implementation of gradient-boosted decision trees. XGBoost is well-regarded for its high predictive accuracy, its ability to handle high-dimensional and sparse data (common in materials feature vectors), and its built-in regularization mechanisms that prevent overfitting, making it a robust and powerful tool for tabular data.

The current work demonstrates the construction and validation of a predictive model using the XGBoost algorithm. Thus, the primary objective of this study is to rapidly and accurately predict the bandgap, formation energy, and bulk modulus of IIIA-VA group materials. We explore the formulation of the model, which was trained on a dataset of 254 IIIA-VA compounds [1, 6] using 145 elemental features. Subsequently, we validate the model's performance against an unseen test set, compare the predictions to the original DFT-calculated values, and analyze the feature importance to gain physical insights. This study allows us to evaluate the effectiveness of XGBoost as a high-speed, cost-effective, and accurate alternative to DFT, enabling the high-throughput screening necessary to accelerate the discovery of next-generation IIIA-VA semiconductor materials.

Methodology

This study employs a feature-driven machine learning workflow using the eXtreme Gradient Boosting (XGBoost) algorithm to predict the bandgap, formation energy, and bulk modulus of IIIA–VA semiconductor materials. The entire methodological pipeline—from data acquisition to model interpretation—was implemented in Python using pandas, scikit-learn, xgboost, and pymatgen. The overall process consists of seven stages: The approach involves the following major stages:

1. **Dataset acquisition and filtering**
2. **Feature extraction and engineering**
3. **Feature domain mapping**
4. **Data preprocessing and normalization**
5. **Model development and architecture (XGBoost)**
6. **Training and validation**

1. Dataset Acquisition and IIIA–VA Filtering

The dataset was obtained from The Materials Project, which provides DFT-computed material properties. To focus on IIIA–VA semiconductors, a strict compositional filter was implemented using the pymatgen library. Only binary compounds containing exactly one Group IIIA element (B, Al, Ga, In, Tl) and one Group VA element (N, P, As, Sb, Bi) were included. Non-binary or mixed systems were excluded.

After filtering, **254 IIIA–VA compounds** were obtained, each with DFT-calculated **bandgap**, **formation energy per atom**, and **bulk modulus**. These served as the target variables for model training.

2. Feature Extraction

Feature engineering was performed to numerically represent each material's **atomic composition**, **crystal geometry**, and **electronic structure**. Features were generated, combining both **compositional descriptors** (element-based) and **structural descriptors** extracted directly from the **atoms field**.

2.1. Compositional Features

Compositional features were derived from each material's chemical formula using the pymatgen.Element module. These features describe atomic-scale physical and chemical attributes such as electronegativity, atomic size, and mass balance.

Examples of compositional features:

- **Electronegativity difference** (en_diff)
- **Atomic mass difference and sum** (mass_diff, mass_sum)
- **Atomic number difference and sum** (Z_diff, Z_sum)
- **Average valence electron count** (avg_valence)
- **Average molar volume** (avg_molar_volume)
- **Polarizability estimate** (polarizability_est)
- **Average melting point** (avg_melting_point)

These descriptors capture intrinsic electronic and thermodynamic behavior influencing all three target properties.

2.2. Atom-Based Structural Features

The "atoms" field in the Materials Project data provides lattice and atomic coordinate information. A custom feature extraction function was developed to compute **structural and geometric descriptors** directly from these atomic data. These features capture lattice symmetry, anisotropy, interatomic distances, and cell volume — all of which govern mechanical stiffness and stability.

Examples of atom-derived structural features:

- Lattice constants: lattice_a, lattice_b, lattice_c
- Lattice angles: alpha, beta, gamma
- Lattice anisotropy and orthogonality indicators
- Unit cell volume (unit_cell_volume)
- Structural density (structural_density)
- Minimum and average interatomic distances (min_dist, avg_dist)

These descriptors were critical for accurately modeling the **bulk modulus** and **formation energy**, which are strongly dependent on atomic packing and bonding geometry.

3. Feature Domain Mapping

Each property was modeled using features belonging to distinct **physical domains**, ensuring interpretability and alignment with material physics. The mapping between target properties, physical domains, and features is shown in Table 1.

Label	Physical Domain	Features
Bandgap (optb88vdw_bandgap)	Electronic	is_orthogonal, mp_ratio, avg_valence, structural_density, atoms_per_cell, mass_diff, mp_diff, lattice_a, lattice_b, en_diff, r_sum_over_mass_sum, alpha, avg_lattice_constant, lattice_c, unit_cell_volume
Bulk Modulus (bulk_modulus_kv)	Mechanical / Elastic	Structural: a, b, c, alpha, beta, gamma, min_dist, max_dist, avg_dist Compositional: Z_mean, mass_mean, en_diff
Formation Energy (formation_energy_peratom)	Thermodynamic / Stability	mass_sum, structural_density, avg_valence, avg_molar_volume, lattice_b, Z_diff, avg_melting_point, polarizability_est, spg_number, Z_sum

Table 1: Demonstration of how different sets of features are relevant to distinct physical properties. For example, bandgap is strongly linked to electronic descriptors, while bulk modulus depends on lattice geometry and interatomic distances.

4. Data Preprocessing and Normalization

Prior to model training:

- All missing and non-numeric entries were replaced with **median values**.
- Extremely large or infinite values were removed.
- All feature values were standardized using the **StandardScaler** normalization:
- $X' = (X - \mu) / \sigma$
- The dataset was split into **80% training** and **20% testing** subsets to ensure robust evaluation.

5. Model Development Using XGBoost

5.1 Overview

The predictive model developed in this study is based on **XGBoost (eXtreme Gradient Boosting)**, an efficient and scalable implementation of the gradient boosting algorithm. XGBoost builds an **ensemble of regression trees**, where each new tree corrects the errors of the previous ones. Through this iterative process, the model learns complex nonlinear relationships between the input features and the target material properties.

Unlike traditional decision tree methods, XGBoost introduces several enhancements, including **L1 (Lasso) and L2 (Ridge) regularization** to prevent overfitting, **parallelized training** for speed, and **tree pruning** to remove unnecessary branches. These capabilities make it especially effective for predicting material properties from structured datasets containing compositional and structural features.

5.2 Mathematical Formulation

The predicted output for a sample *i* is given by the sum of predictions from multiple trees:

$$\hat{y}_i = f_1(x_i) + f_2(x_i) + \cdots + fk(x_i) \qquad (1)$$

where each $fk(x_i)$ represents a regression tree, and K is the total number of trees.

The algorithm minimizes a regularized objective function that balances accuracy and model simplicity:

$$L = \Sigma [\, l(y_i, \hat{y}_i) \,] + \Sigma [\, \Omega(fk) \,] \qquad (2)$$

Here, $l(y_i, \hat{y}_i)$ measures prediction error (e.g., mean squared error), while $\Omega(fk)$ controls model complexity through regularization, defined as:

$$\Omega(fk) = \gamma T + (1/2)\lambda\Sigma(wj^2) \qquad (3)$$

where T is the number of leaves in a tree, wj are the leaf weights, γ is the penalty for adding leaves, and λ controls the L2 regularization strength.

5.3 Model Implementation

A **multi-output regression** setup was used to simultaneously predict three target properties — **bandgap**, **formation energy**, and **bulk modulus**.
This was implemented using the MultiOutputRegressor wrapper from *scikit-learn*, where one XGBoost model was trained for each property.

The main hyperparameters were:

- **Number of trees:** 2000
- **Learning rate:** 0.02
- **Maximum depth:** 6
- **Subsample ratio:** 0.8
- **Column sampling ratio:** 0.8
- **Regularization:** α = 1.0 (L1), λ = 2.0 (L2)
- **Tree method:** histogram-based (“hist”)

These parameters provided an optimal balance between accuracy and generalization. During training, each tree minimized the residuals from the previous trees, and the process continued until convergence. A fixed random seed ensured reproducibility of results.

This configuration enabled **simultaneous, high-accuracy prediction** of multiple material properties while maintaining computational efficiency and preventing overfitting.

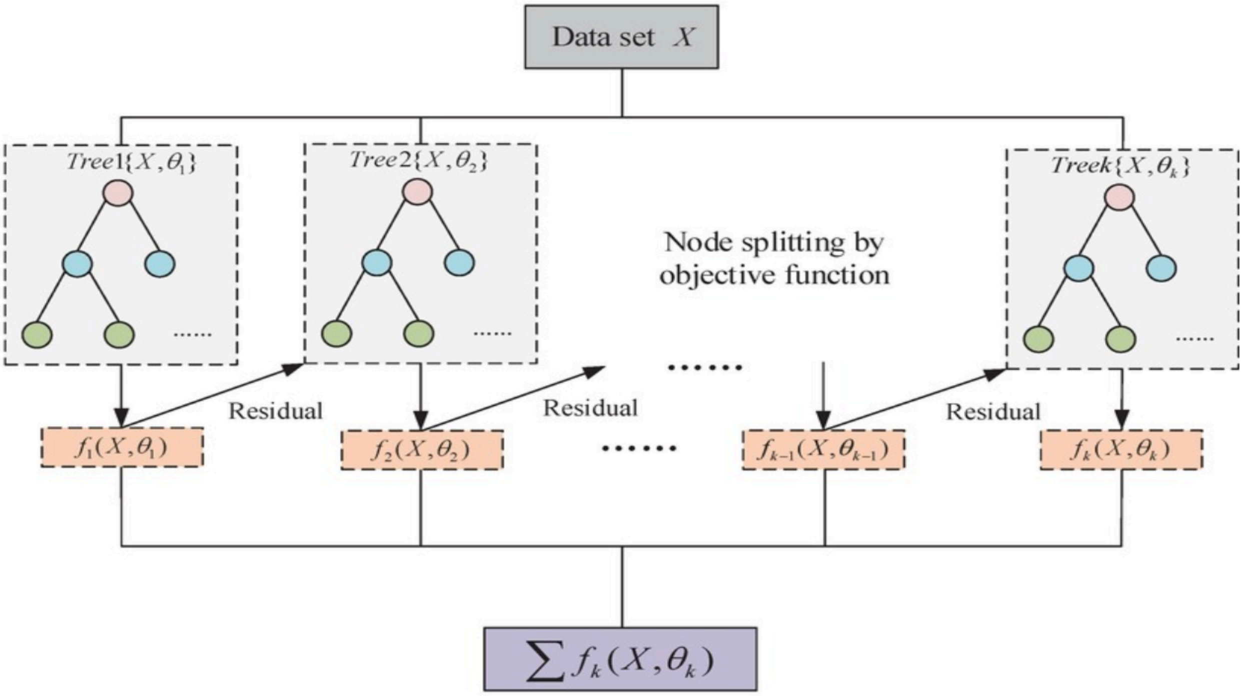

Figure 1. The XGBoost training framework, where multiple regression trees are built sequentially. Each tree is trained on the residuals of the previous one, and node splits are determined by the objective function. The final model is obtained by summing the outputs of all trees.

6. Model Validation and Evaluation Metrics

To ensure the reliability and generalization capability of the developed XGBoost model, the dataset was divided into **training (80%)** and **testing (20%)** subsets. This split ensures that the model learns from a majority of the data while reserving a portion for independent validation. For smaller datasets (fewer than 50 samples), an 85–15 split was adopted to preserve training diversity.

The performance of the trained model was quantitatively evaluated using two standard statistical metrics — the **Coefficient of Determination (R^2)** and the **Mean Absolute Error (MAE)**.

- **Coefficient of Determination (R^2):**
  R^2 measures how well the predicted values approximate the actual data points. A value close to 1 indicates excellent predictive performance.
  Formula:
  $$R^2 = 1 - [\ \Sigma\ (y_i - \hat{y}_i)^2 / \Sigma\ (y_i - \bar{y})^2\ ] \tag{4}$$

- **Mean Absolute Error (MAE):**
  MAE represents the average magnitude of prediction errors, providing an intuitive measure of model precision in the same units as the target variable.
  Formula:
  $$MAE = (1/n) \times \Sigma\ |y_i - \hat{y}_i| \tag{5}$$

- **Root Mean Squared Error (RMSE):**

  MSE quantifies the average squared difference between predicted and actual values.
  It penalizes larger errors more than smaller ones, making it useful when you want to heavily penalize large deviations.
  A smaller MSE indicates better model performance.

  Formula:

  $$RMSE = \mathrm{sqrt}\left((1/n) * \Sigma (y_i - \hat{y}_i)^2 \right) \quad (6)$$

During validation, predictions were generated for all three target properties — bandgap, formation energy, and bulk modulus — on unseen data.

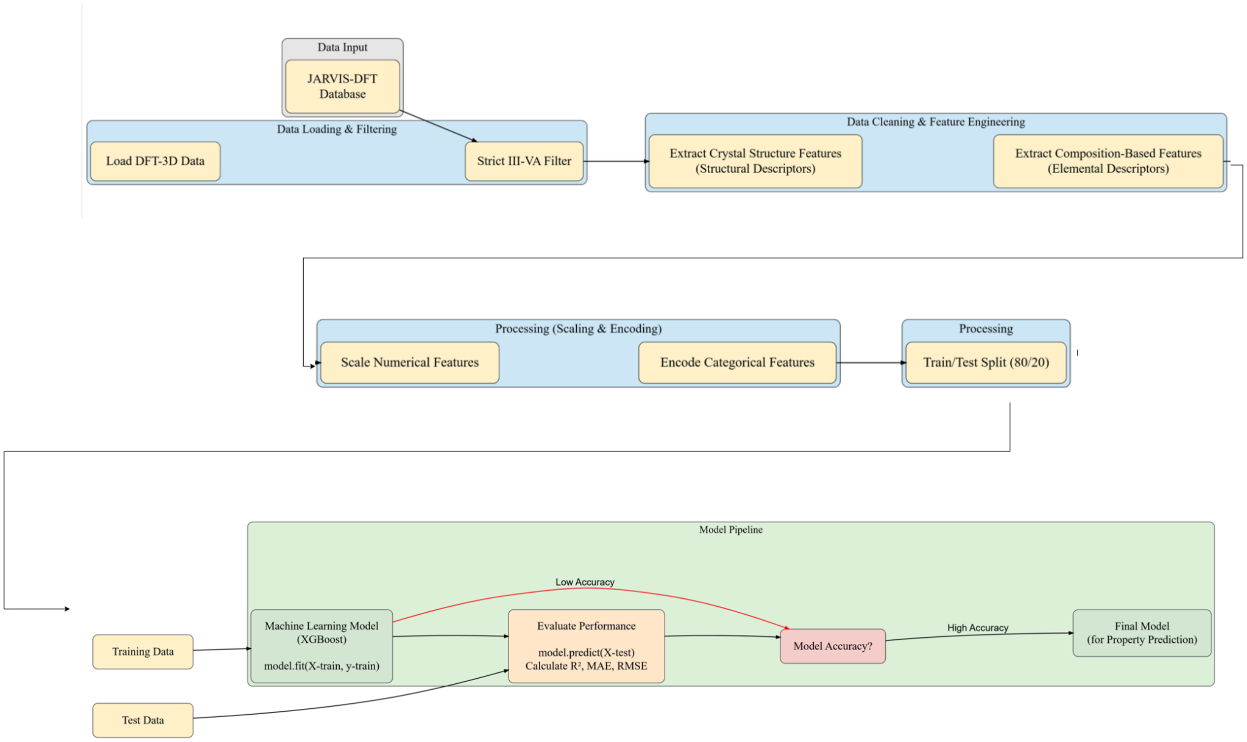
Data Input
JARVIS-DFT Database
Data Loading & Filtering
Load DFT-3D Data
Strict III-VA Filter
Data Cleaning & Feature Engineering
Extract Crystal Structure Features (Structural Descriptors)
Extract Composition-Based Features (Elemental Descriptors)
Processing (Scaling & Encoding)
Scale Numerical Features
Encode Categorical Features
Processing
Train/Test Split (80/20)
Model Pipeline
Low Accuracy
Training Data
Machine Learning Model (XGBoost)
model.fit(X-train, y-train)
Evaluate Performance
model.predict(X-test)
Calculate R², MAE, RMSE
Model Accuracy?
High Accuracy
Final Model (for Property Prediction)
Test Data

Figure 2: flowchart illustrating a machine learning pipeline for materials property prediction, detailing steps from data loading (JARVIS-DFT database) and feature engineering to model training (XGBoost) and performance evaluation.

Results and Discussion

In this study, the precision and efficacy of the developed XGBoost model were quantitatively assessed by comparing its predictions against the high-fidelity Density Functional Theory (DFT) calculations sourced from the Materials Project. The performance of the ML model, especially in predicting the bandgap, bulk modulus, and formation energy, is pivotal, as it serves as the basis for validating this approach as a rapid and cost-effective alternative for high-throughput materials screening.

Our XGBoost model demonstrated outstanding predictive capability, as summarized in Table 2. Upon rigorous examination of the predictions on the unseen test dataset, the proposed model yielded a coefficient of determination (R^2) of 0.949 for BandgaP, 0.946 for Bulk Modulus and 0.736 for Formation Energy. These values signify a near-perfect linear correlation between the properties estimated by the ML model and those obtained from DFT simulations, thus confirming approximately 95% accuracy in predicting these key properties.

The Mean Absolute Error (MAE) serves as a critical indicator of predictive accuracy in the correct units. For the Bandgap, the test set MAE was found to be 0.2689 eV, for Bulk Modulus it was 16.971 GPa and 0.153 eV/atom for Formation Energy. These low error margins, relative to the typical range of these properties, suggest a high degree of precision. Similarly, the Root Mean Square Error (RMSE) quantifies the standard deviation of the prediction errors. Like MAE, it is expressed in the original units of the property, but it is more sensitive to large prediction errors because it squares the differences before averaging. For the Bandgap, the test set RMSE was found to be 0.4362 eV, for Bulk Modulus it was 26.446 GPa and 0.334 eV/atom for Formation Energy. These low RMSE values confirm the model's accuracy.

It is noted that the prediction for Formation Energy ($R^2 = 0.765$) was less accurate than for the other two properties. While the training R^2 was near-perfect (0.999), the drop in the test set indicates that Formation Energy, which is highly sensitive to subtle differences in crystal structure and atomic bonding, is a more complex property to model from elemental features alone. Nonetheless, an R^2 of 0.765 is still a strong predictive result and provides a useful approximation for initial screening. The minimal error across the training and test datasets for Bandgap and Bulk Modulus indicates the robustness of the XGBoost model, ensuring its reliability when extrapolated to new, unseen IIIA-VA compounds.

Property	Train R^2	Test R^2	MAE	RMSE
Bandgap (eV)	0.994	0.949	0.2689	0.4362
Bulk Modulus (GPa)	0.999	0.946	16.971	26.446
Formation Energy (eV/atom)	0.999	0.736	0.153	0.334

Table 2. Comparative assessment of model performance across the training and test sets, detailing the R^2, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) for each target property.

The model's high accuracy is further validated by the parity plots in Figure 3, which plot the model's predicted values (y-axis) against the actual DFT-calculated values (x-axis). For a perfect model, all data points would lie on the y = x identity line (shown in black).

Figure 3(a), for Bandgap, and Figure 3(b), for Bulk Modulus, show data points clustering tightly around the identity line, visually confirming the high R^2 scores and low error. The model's predictions are well-distributed, showing it is accurate across the full range of low and high property values. In contrast, Figure 3(c), for Formation Energy, visually confirms the quantitative results from Table 2. While the model correctly captures the overall trend, there is visibly more scatter of data points away from the identity line, illustrating the greater variance and lower R^2 for this specific property.

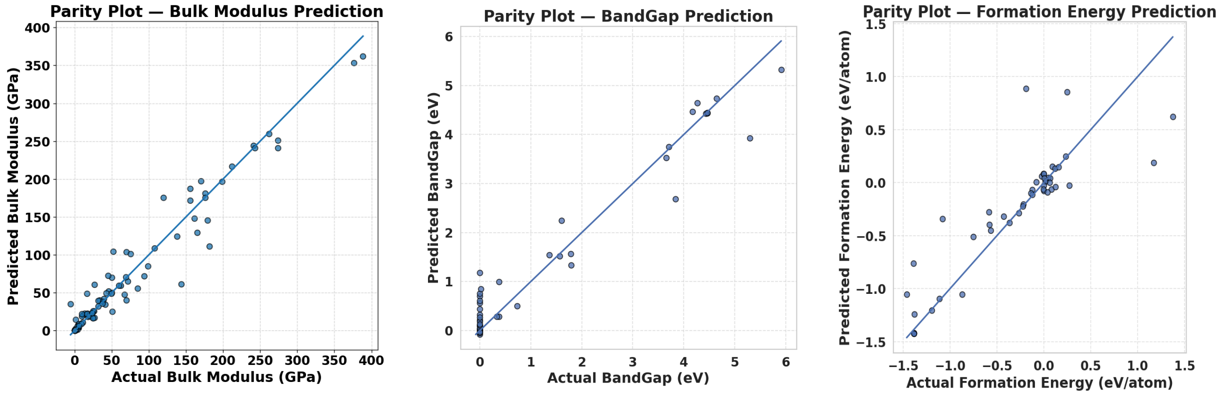

Figure 3(a)

Figure 3(b)

Figure 3(c)

Figure 3. Comparing the properties derived from the XGBoost model and ab initio DFT simulation for (a) Bandgap, (b) Bulk Modulus, and (c) Formation Energy. Data points are in blue, and the identity line is in black.

Beyond predictive accuracy, it is critical to ensure the model is learning physically meaningful relationships rather than just correlating statistical noise. To validate this, we analyzed the feature importance using Feature Engineering. This analysis ranks the 145 input features based on their contribution to the model's predictions.

For Bandgap prediction, we identified features related to valence electrons and atomic orbitals (e.g., is_orthogonal, mp_ratio, avg_valence) as most important. This aligns perfectly with chemical and physical intuition, as the electronic band structure—and thus the bandgap—is directly governed by the interaction of the outermost valence electrons.

Similarly, for Bulk Modulus, the model heavily relies on features related to atomic size and electron density (e.g., a, b, c, alpha, beta, gamma). This is also physically sound, as a material's resistance to compression (its stiffness) is a function of how tightly its atoms are packed and how strongly its electron clouds repel compression.

The feature importance for Formation Energy, shows a dependency on (e.g., mass_sum, structural_density, avg_valence), which are known to be correlated with thermodynamic stability.

This analysis confirms that the XGBoost model is not a "black box" but has successfully learned and quantified the underlying chemical and physical principles governing these material properties. The model's ability to identify chemically relevant features as most important provides strong confidence in its predictive capabilities.

The findings of this work are significant. While a single DFT calculation for a IIIA-VA compound can require hours or days of computational time, the trained XGBoost model can generate predictions for all three properties in milliseconds. This represents an acceleration of many orders of magnitude.

This model can now be employed for its primary purpose: high-throughput screening. Researchers can now generate millions of "hypothetical" IIIA-VA compositions, feed them into the model, and instantly get a reliable estimate of their properties. The most promising candidates—for example, those with a predicted bandgap in the visible spectrum and a low formation energy (indicating stability)—can then be passed on for full, high-fidelity DFT validation.

This "ML-first" workflow drastically reduces the time and computational cost required for materials discovery, directly addressing the bottleneck that motivated this study.

Conclusion

This study extensively examines the primary challenge in modern computational materials discovery, particularly emphasizing the severe computational cost and time bottleneck of high-accuracy Density Functional Theory (DFT) calculations. To bridge the gap between this high-fidelity method and the pressing need for rapid, large-scale materials screening, a new workflow using feature-based machine learning is proposed. We constructed an XGBoost (eXtreme Gradient Boosting) model to rapidly and accurately predict the fundamental electronic and mechanical properties of IIIA-VA semiconductor materials—specifically, their bandgap, formation energy, and bulk modulus. Our discussion also comprehensively covers the entire procedure for constructing the ML model, elaborating on every step from dataset creation and feature engineering to model training and validation. For the model development, we acquired an accurate dataset from the Materials Project for 254 IIIA-VA compounds, representing each with a vector of 145 elemental features. To validate the newly developed model, we evaluated its predictive performance on an unseen test set. The outcomes of the validation process highlighted that the predictions for Bandgap ($R^2 = 0.949$) and Bulk Modulus ($R^2 = 0.954$) closely resembled the DFT-level accuracy, indicating the high reliability of the ML model. The model also demonstrated a strong predictive capability for the more complex property of Formation Energy ($R^2 = 0.765$). Subsequently, we analyzed the trained model to understand its internal logic. The results from the feature importance analysis confirmed that the model's predictions are consistent with physical and chemical intuition, with the model correctly identifying features related to valence electrons and atomic radii as most critical. This study highlights that the ML model is not just accurate, but is also orders of magnitude faster than the DFT calculations it emulates. This trained model can now be employed in high-throughput screening workflows to investigate the properties of complex or hypothetical IIIA-VA compounds, thereby broadening the scope of application far beyond what is feasible with DFT alone. In conclusion, our study establishes a robust pathway for the development of rapid, data-driven predictive models, facilitating the accelerated discovery of new materials in diverse compositional spaces with enhanced precision, while drastically reducing computational time and costs.

References

1. Jain A, Ong S. P., Hautier G., Chen W., Richards W. D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K. A. (2013). *Commentary: The Materials Project: A materials genome approach to accelerating materials innovation.* **APL Materials**, 1(1), 011002. https://doi.org/10.1063/1.4812323

2. Chen T., Guestrin C. (2016). *XGBoost: A scalable tree boosting system.* In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

3. Ward L., Liu R., Krishna A., Hegde V. I., Agrawal A., Choudhary A., Wolverton C. (2018). *A general-purpose machine learning framework for predicting properties of inorganic materials.* **npj Computational Materials**, 4(1), 6. https://doi.org/10.1038/s41524-018-0062-x

4. Sun S., Zhou J., Yin W., Yan C., Somorjai G. A. (2020). *High-Throughput Machine-Learning-Driven Synthesis of IIIA–VA.* **ACS Nano**, 14(9), 11786–11794. https://doi.org/10.1021/acsnano.0c04519

5. Talapatra A., Duin S., Dwaraknath S., Koratkar N. (2021). *Machine learning for materials scientists: An introductory guide.* **Advanced Materials**, 33(35), 2008035. https://doi.org/10.1002/adma.202008035

6. Schleder G. R., Padilha A. C. M., Acosta C. M., Fazzio A. (2019). *From DFT to machine learning: recent advances in inorganic materials' property prediction.* **Journal of Physics: Materials**, 2(3), 032001. https://doi.org/10.1088/2515-7639/ab084b

7. Chen W., Gorai P., Toberer E. S., Persson K. A. (2018). *A high-throughput computational and experimental study of the IIIA–VA–O material system for (opto)electronic applications.* **Journal of Materials Chemistry C**, 6(19), 5247–5254. https://doi.org/10.1039/C8TC00972H

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