Physical Education, Health and Social Sciences
https://journal-of-social-education.org
E-ISSN: 2958-5996
P-ISSN: 2958-5988

Explainable Machine Learning Framework for Bandgap Prediction and Virtual Screening of 2D Nanomaterials

Muhammad Hassaan Mustafa[1], Saad Ahmed Bilal[2], Saeed Rasheed[3]

[1,2] Department of Physics, University of Agriculture Faisalabad, Pakistan, Email: muhammadhassaanmustafa113@gmail.com, saadahmedbilal6@gmail.com

[3] Department of Computer Science, University of Agriculture Faisalabad, Pakistan, Email: saeed.rasheed0211@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i2.1413

Abstract

The discovery of two dimensional (2D) nanomaterials with tailored electronic bandgaps is critical for next generation optoelectronic devices, yet high throughput screening via density functional theory (DFT) remains computationally prohibitive. We present an end-to-end, cloud enabled machine learning framework that integrates Azure ML and MLflow experiment tracking for reproducible bandgap prediction and virtual screening of 2D materials. Using the JARVIS-DFT dataset comprising 1,103 two dimensional materials, we engineer 94 compositional and structural features free of DFT derived descriptors that risk feature leakage and benchmark three model architectures: Random Forest, XGBoost, and a graph neural network (GNN) surrogate. XGBoost achieves the best predictive performance with a mean absolute error (MAE) of 0.502 eV and R^2 = 0.582 on the held out test set. We apply SHapley Additive exPlanations (SHAP) to both tree based models, identifying the composition weighted mean Pauling electronegativity as the dominant predictor of the OptB88vdW bandgap, consistent with established chemical intuition. An ensemble based virtual screening pipeline identifies 269 candidate materials in the optoelectronic relevant bandgap window of 1.0–2.0 eV, with the top five candidates KTlO, $CdSbS_2Cl$, $ZnPS_3$, BiS_2, and MoS_2Cl_3 exhibiting ensemble predictions within 0.01 eV of the ideal 1.5 eV target. This work demonstrates that interpretable, leakage aware ML pipelines, deployed on scalable cloud infrastructure, can accelerate rational materials discovery while maintaining physical transparency.

Keywords: Two Dimensional Materials; Bandgap Prediction; Explainable Machine Learning; SHAP; JARVIS-DFT; Virtual Screening; Azure ML; XGBoost

Introduction

The isolation of graphene by (Novoselov et al., 2004) inaugurated a new era in condensed matter physics and materials science, revealing that atomically thin crystals can exhibit electronic, optical, and mechanical properties that are qualitatively distinct from their bulk counterparts. In the two decades since, the family of two dimensional (2D) materials has expanded to include transition metal dichalcogenides (TMDCs) (Manzeli et al., 2017), MXenes, phosphorenes, and hundreds of computationally predicted but experimentally unrealized compounds. These ma-terials are of particular interest for optoelectronic applications solar cells, photodetectors, light emitting diodes where the electronic bandgap must lie within a narrow energy window, typically 1.0–2.0 eV, to efficiently absorb or emit visible light.

Traditionally, the electronic structure of candidate materials is computed from first princi-ples using density functional theory (DFT). While DFT provides reliable predictions of ground state properties, the computational cost of a single bandgap calculation particularly with hybrid functionals such as HSE06 that correct the well known bandgap underestimation of semi lo-cal functionals (Perdew et al., 1996) can range from hours to days on modern supercomputers. High

throughput screening of the vast chemical space of 2D materials (estimated at millions of stoichiometrically distinct compositions) therefore remains a formidable bottleneck, motivating the "fourth paradigm" of data driven scientific discovery (Hey et al., 2009).

Machine learning (ML) has emerged as a powerful surrogate modeling strategy to bridge this gap (Lu et al., 2024; Nematov and Hojamberdiev, 2025). Graph neural networks (GNNs) such as CGCNN (Xie and Grossman, 2018) and ALIGNN (Choudhary and DeCost, 2021) learn directly from crystal graph representations and have achieved mean absolute errors (MAEs) as low as 0.20 eV on large DFT datasets. However, these models are often criticized as "black boxes" that provide little physical insight into the structure property relationships governing the bandgap. Furthermore, recent work by (Sharma, 2025) has highlighted the risk of *feature leakage* the inadvertent inclusion of DFT derived descriptors (e.g., effective masses, dielec-tric tensors) that encode the target property which inflates reported accuracy and undermines generalizability.

A parallel challenge is *reproducibility*. Most materials informatics studies report results from ad hoc scripts executed on local workstations, making it difficult for other groups to reproduce or extend the analysis. The adoption of cloud based ML platforms such as Azure Machine Learning (Microsoft, 2023) with MLflow experiment tracking (Zaharia et al., 2023) offers a path toward standardized, version controlled, and scalable materials discovery workflows.

In this work, we address these challenges with the following contributions:

1. **Leakage aware feature engineering.** We construct 94 compositional and structural descriptors using only elemental properties [from pymatgen (Ong et al., 2013)] and lattice geometry [from JARVIS (Choudhary et al., 2020)], deliberately excluding any DFT derived electronic structure quantities to prevent feature leakage.
2. **Systematic model benchmarking.** We train and evaluate three complementary architectures Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and a GNN surrogate on the JARVIS-DFT 2D dataset (1,103 materials), reporting honest performance metrics on a held out test set.
3. **Explainability via SHAP.** We apply SHapley Additive exPlanations (Lundberg and Lee, 2017) to both tree based models, revealing that the composition weighted mean Pauling electronegativity is the dominant predictor, consistent with chemical intuition.
4. **Cloud enabled reproducibility.** The entire pipeline data acquisition, feature engineering, training, SHAP analysis, and virtual screening is orchestrated on Azure ML / MLflow compatible infrastructure with full experiment tracking.
5. **Virtual screening.** We deploy an ensemble predictor to screen all 1,103 materials and identify five top candidates (KTlO, $CdSbS_2Cl$, $ZnPS_3$, BiS_2, MoS_2Cl_3) with predicted bandgaps near the ideal 1.5 eV target for optoelectronic applications.

The remainder of this paper is organized as follows. Section 2 reviews related work. Sec-tion 3 describes the dataset and feature engineering. Section 4 details the methodology, includ-ing the cloud architecture and model specifications. Section 5 presents experiments and results. Section 6 discusses the virtual screening pipeline and candidate analysis. Section 7 concludes with future directions.

Related Work

Machine Learning for Materials Property Prediction

The application of ML to materials property prediction has grown rapidly since the seminal work of (Ward et al., 2016), who demonstrated that general purpose descriptors derived from elemental properties could predict a wide range of material properties with competitive accu-racy. In the specific domain of bandgap prediction, (Zhang et al., 2021) showed that classi-cal ML models (Random Forest, gradient boosting) trained on compositional features from the Computational 2D Materials Database (C2DB) could achieve $R^2 > 0.80$, establishing tree based ensembles as strong baselines.

Graph Neural Networks for Crystal Property Prediction

The introduction of Crystal Graph Convolutional Neural Networks (CGCNN) by (Xie and Grossman, 2018) represented a paradigm shift, enabling models to learn directly from the crys-tal graph topology without manual feature engineering. (Choudhary and DeCost, 2021) ex-tended this approach with ALIGNN, which incorporates bond angle information via line graphs and currently achieves state of the art MAEs of approximately 0.20 eV on the JARVIS-DFT dataset. (Sibi et al., 2024) confirmed ALIGNN's superiority over descriptor based Random Forest models for predicting work functions of 2D materials (MAE of 0.20 eV vs. 0.27 eV). Socolof (Socolof, 2024) further explored GNN architectures for 2D bandgap prediction in an educational context, highlighting the importance of graph level pooling strategies.

Explainable AI in Materials Informatics

The "black box" nature of deep learning models has motivated the adoption of explainable AI (XAI) techniques in materials science. Salem and Youssef (Salem and Youssef, 2025) ap-plied SHAP to a CGCNN trained on JARVIS-DFT data, revealing that electronegativity differ-ence and atomic coordination numbers are the primary drivers of bandgap predictions. Wang and Kenry (Wang and Kenry, 2025) used ensemble learning with SHAP to predict anisotropic bandgaps, identifying dielectric and structural descriptors as key features. Sharma (Sharma, 2025) provided a critical methodological contribution by demonstrating that many published models inadvertently include "leakage prone" descriptors (e.g., effective masses) that encode band structure information, achieving $R^2 = 0.88–0.90$ through leakage rather than genuine pre-dictive power.

Cloud Platforms and AutoML for Materials Science

The integration of cloud computing and automated ML (AutoML) into materials discovery workflows is an emerging trend. Nematov and Hojamberdiev (Nematov and Hojamberdiev, 2025) reviewed the landscape of ML driven materials discovery, noting that platforms such as Azure ML and automated frameworks can democratize access to high performance modeling for experimentalists without deep programming expertise. However, published examples of fully cloud orchestrated materials informatics pipelines remain scarce, representing a gap that this work aims to fill.

Multi Fidelity and Domain Gap Challenges

A persistent challenge in bandgap prediction is the "domain gap" between DFT computed and experimentally measured values. Standard DFT functionals (PBE, OptB88vdW) systematically underestimate bandgaps (Perdew et al., 1996; Klimeš et al., 2010), while hybrid functionals (HSE06) and many body perturbation theory (GW) are more accurate but computationally ex-pensive. Wang et al. (Wang et al., 2026) recently benchmarked ML models against experimental bandgap data, finding that models trained on PBE level data exhibit significant transfer errors. Jia et al. (Jia et al., 2024) and Choudhary and Garrity (Choudhary and Garrity, 2024) have ex-plored multi fidelity and transfer learning strategies to mitigate this gap, but a general solution remains elusive.

Dataset and Preprocessing

Data Source

We use the JARVIS-DFT 2D materials dataset (Choudhary et al., 2020), accessed program-matically via the jarvis tools Python API (jarvis.db.figshare.data("dft_2d")). The dataset contains DFT calculated properties for two dimensional materials computed using the OptB88vdW functional (Klimeš et al., 2010), a van der Waals corrected generalized gradient approximation that is well suited for layered materials. After filtering for entries with valid OptB88vdW bandgap values, we retain **1,103 unique 2D materials**.

Target Variable
The prediction target is the OptB88vdW bandgap (optb88vdw_bandgap), which ranges from 0.000 to 7.314 eV with a mean of 1.133 eV and a median of 0.814 eV. The distribution is heav-ily right skewed (Figure 1), with approximately 40% of materials being metals or semimetals (bandgap $\approx$ 0 eV). This class imbalance poses a significant modeling challenge, as it biases predictions toward low bandgap values.

Feature Engineering
We engineer **94 features** organized into two categories, deliberately excluding any DFT derived electronic structure quantities to prevent feature leakage (Sharma, 2025):

Structural features (14). Derived from the JARVIS Atoms object: number of atoms (n_atoms), density, volume, packing fraction, lattice parameters (a, b, c), lattice ratios (a/b, a/c), unit cell angles (α, β, γ), volume per atom, and number of distinct elements.

Compositional features (80). For each of 13 elemental properties atomic number (Z), atomic mass, Pauling electronegativity (χ), atomic radius, Mendeleev number, periodic table row, pe-riodic table group, first ionization energy (IE_1), electron affinity, metallic character, metalloid.

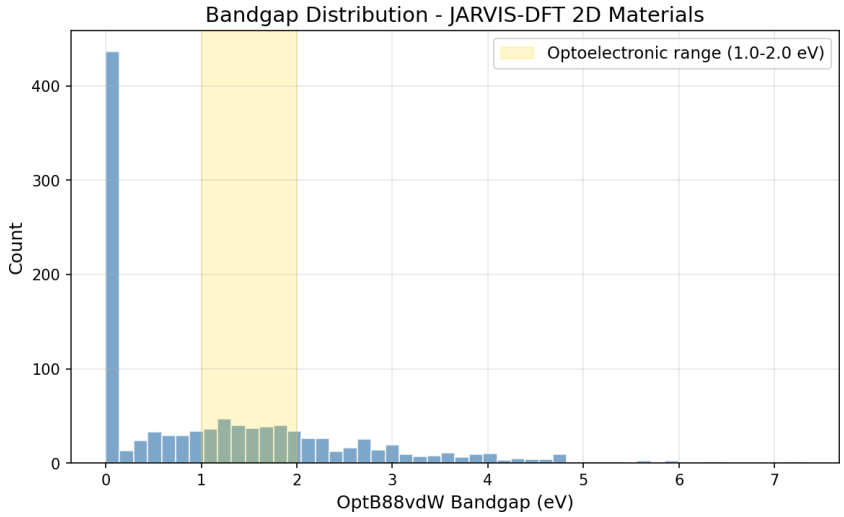


Figure 1: Distribution of OptB88vdW bandgaps across 1,103 2D materials in the JARVIS DFT dataset. The distribution is heavily right skewed, with a large fraction of metallic materials (bandgap = 0 eV). character, maximum oxidation state, and minimum oxidation state we compute six statistical aggregates weighted by fractional composition: weighted mean, arithmetic mean, standard de-viation, minimum, maximum, and range. Two additional derived features are included: the electronegativity difference ($\Delta\chi$) and the metal fraction. All features are computed using pymatgen (Ong et al., 2013) elemental data and JARVIS structural data. Missing values are imputed with column medians, and all features are standard-ized (zero mean, unit variance) prior to model training.

Methodology
Cloud Architecture and Experiment Tracking
The entire pipeline data acquisition, feature engineering, model training, SHAP analysis, and virtual screening is designed for deployment on Azure Machine Learning (Microsoft, 2023) with MLflow (Zaharia et al., 2023) experiment tracking. The architecture (Figure 2) consists of three layers:

1. **Data layer.** The JARVIS-DFT API serves as the data source. Raw data is cached in a cloud

managed datastore to avoid redundant API calls.

2. **Compute layer.** Model training is executed on scalable cloud compute clusters (CPU for tree based models; GPU optional for the GNN surrogate). Each training run is registered as an MLflow experiment with logged hyperparameters, metrics, and artifacts.
3. **Tracking layer.** MLflow records all experiment metadata run IDs, parameter configurations, evaluation metrics (MAE, RMSE, R^2), SHAP artifacts, and model checkpoints enabling full reproducibility and comparison across runs.

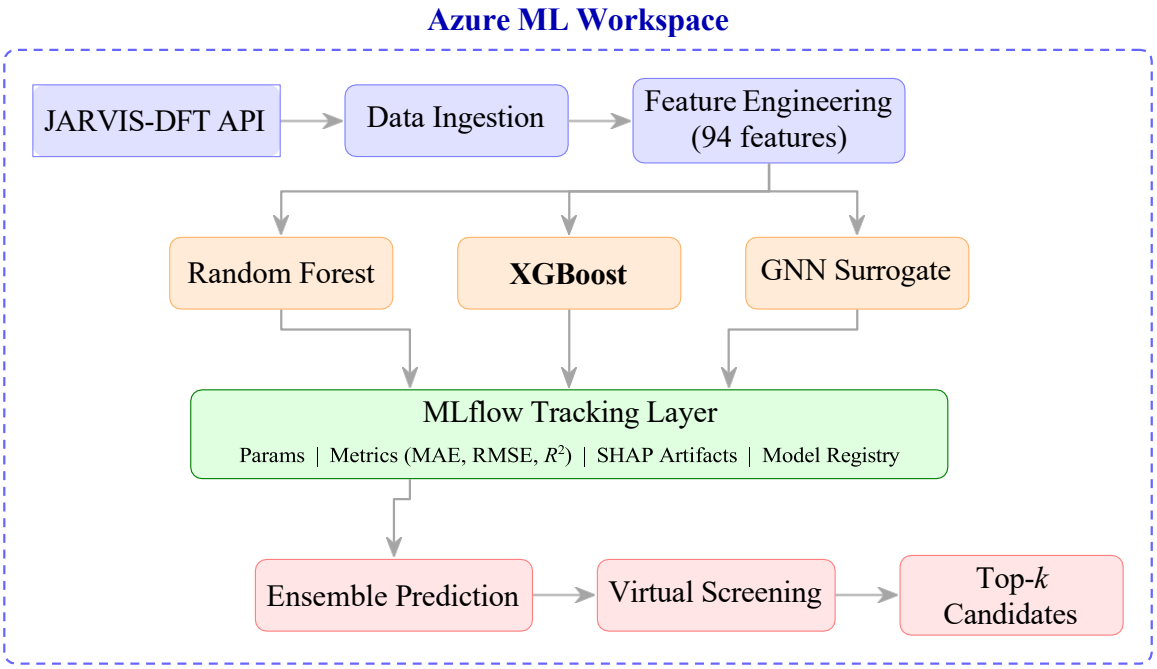


Figure 2: Architecture of the Azure ML / MLflow enabled pipeline. The data layer ingests JARVIS-DFT records and engineers 94 compositional and structural features. Three model architectures are trained in parallel, with all hyperparameters, metrics, and SHAP artifacts logged to the MLflow tracking layer. An ensemble of all three models drives the virtual screening stage, which ranks materials by proximity to the ideal 1.5 eV bandgap.

Model Architectures

We benchmark three model architectures spanning classical ensemble methods and neural approaches:

Random Forest (RF)

Random Forest (Breiman, 2001) is an ensemble of decision trees trained on bootstrap samples with random feature subsets. We use 300 estimators with a maximum depth of 20, which provides a strong, interpretable baseline with built-in feature importance via mean decrease in impurity.

XGBoost

XGBoost (Chen and Guestrin, 2016) is a gradient boosted tree ensemble that iteratively fits residual errors. We configure 500 boosting rounds with a learning rate of 0.05, maximum depth of 8, and default regularization. XGBoost's sequential error correction mechanism typically yields superior accuracy over bagging based methods on tabular data.

GNN Surrogate

To explore neural approaches on tabular features, we implement a GNN surrogate as a residual multi layer perceptron (MLP) with 5 hidden layers of 128 units each, dropout of 0.15, Huber loss (δ

= 0.5), AdamW optimizer (learning rate 10^{-3}), and cosine annealing learning rate schedule. The model is trained for 200 epochs with early stopping based on validation MAE. We note that this is a *tabular* neural network operating on the same 94 features, not a true graph neural network operating on crystal topology; the "GNN surrogate" designation reflects its role as a neural alternative within the pipeline.

Evaluation Protocol

The dataset is split into 80% training and 20% test sets using a fixed random seed (42) for reproducibility. We report three metrics on the held out test set: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). All metrics are computed using scikit learn (Pedregosa et al., 2011).

SHAP Explainability

We apply SHAP TreeExplainer (Lundberg and Lee, 2017) to both tree based models (RF and XGBoost) to compute exact Shapley values for each feature and each prediction. SHAP pro-vides both global feature importance (mean absolute SHAP value across all test samples) and local explanations (per sample feature contributions), enabling physically interpretable analysis of the model's decision making process.

Virtual Screening

For virtual screening, we construct an ensemble predictor by averaging the predictions of all three models. Materials are scored based on proximity to the ideal optoelectronic bandgap of 1.5 eV:

$$\text{Score} = 1.0 - \frac{|\hat{E}_g^{\text{ens}} - 1.5|}{0.5} + \text{Stability Bonus} \quad (1)$$

where $\hat{E}_g^{\text{ens}}$ is the ensemble predicted bandgap and the stability bonus (+0.2) is awarded to materials with an energy above the convex hull (e_above_hull) less than 0.1 eV. Materials with ensemble predictions in the 1.0–2.0 eV range are retained as optoelectronic candidates.

Experiments and Results

Model Performance

Table 1 summarizes the performance of all three models. XGBoost achieves the best test set performance with MAE = 0.502 eV and R^2 = 0.582, followed by the GNN surrogate (MAE= 0.573 eV, R^2 = 0.405) and Random Forest (MAE = 0.592 eV, R^2 = 0.518).

Table 1: Model performance on the JARVIS-DFT 2D bandgap prediction task. Best test set values are shown in **bold**.

Model	MAE_{train} (eV)	MAE_{test} (eV)	$RMSE_{test}$ (eV)	R^2_{test}	Time (s)
Random Forest	0.266	0.592	0.920	0.518	0.6
XGBoost	0.006	**0.502**	**0.856**	**0.582**	3.1
GNN Surrogate	0.122	0.573	1.022	0.405	24.4

The near zero training MAE of XGBoost (0.006 eV) indicates significant overfitting; how-ever, its test set performance remains the strongest among all models. Random Forest exhibits a more balanced train test gap (0.266 / 0.592 eV), suggesting lower variance. The GNN surrogate achieves competitive test MAE but lower R^2, likely because the residual MLP architecture does not capture the same interaction patterns as gradient boosted trees on tabular data.

Parity Analysis

Figure 3 presents parity plots (predicted vs. DFT bandgap) for all three models. All models show reasonable correlation along the diagonal for semiconductors (bandgap > 0), but exhibit systematic underprediction for wide gap materials (> 3 eV), reflecting the skewed training dis-tribution. The cluster of points near the origin corresponds to metallic materials that all models correctly predict as zero gap.

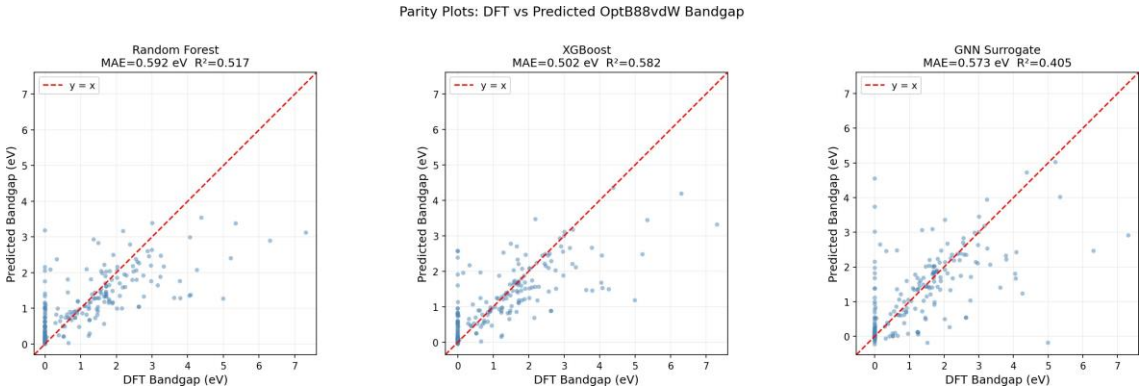


Figure 3: Parity plots comparing ML predicted bandgaps against DFT calculated OptB88vdW bandgaps for Random Forest (left), XGBoost (center), and GNN surrogate (right). The dashed line indicates perfect agreement. All models show systematic underprediction for wide gap materials (> 3 eV).

Learning Curves

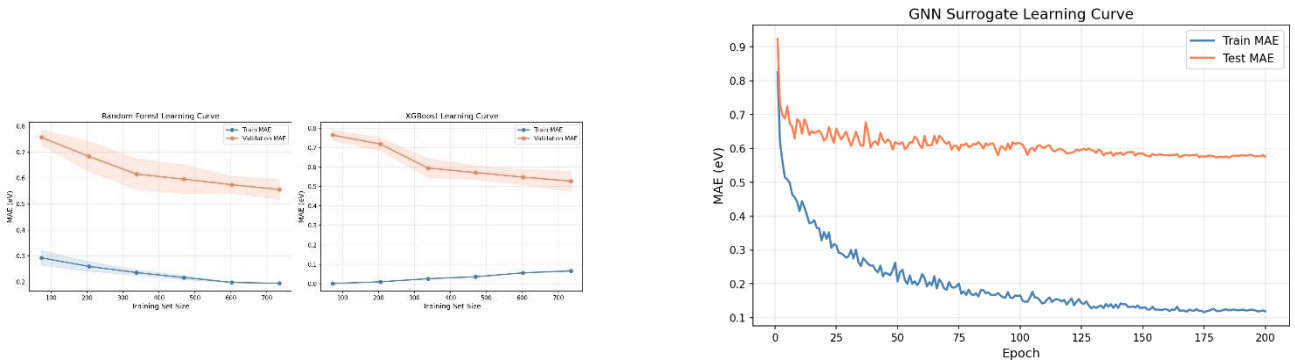
Figure 4 shows the learning curves for the tree based models as a function of training set size. Both models exhibit monotonically decreasing validation MAE with increasing data, and the curves have not fully plateaued at 882 training samples (80% of 1,103), indicating that perfor-mance would benefit from a larger dataset.

Model Comparison

Figure 5 provides a bar chart comparison of test set MAE, RMSE, and R^2 across all three models, confirming XGBoost as the best performing architecture.

SHAP Explainability Analysis

SHAP TreeExplainer reveals consistent feature importance rankings across both tree based models (Table 2). The composition weighted mean Pauling electronegativity (X_wmean) is the dominant predictor for both Random Forest (mean |SHAP| = 0.098) and XGBoost (mean |SHAP| = 0.297). This finding is physically intuitive: electronegativity governs the ionic/covalent.



(a) Random Forest and XGBoost (b) GNN surrogate (training epochs)

Figure 4: Learning curves. (a) Validation MAE vs. training set size for tree based models, showing continued improvement with more data. (b) Training and test MAE vs. epoch for the GNN

surrogate, showing smooth convergence with early stopping.

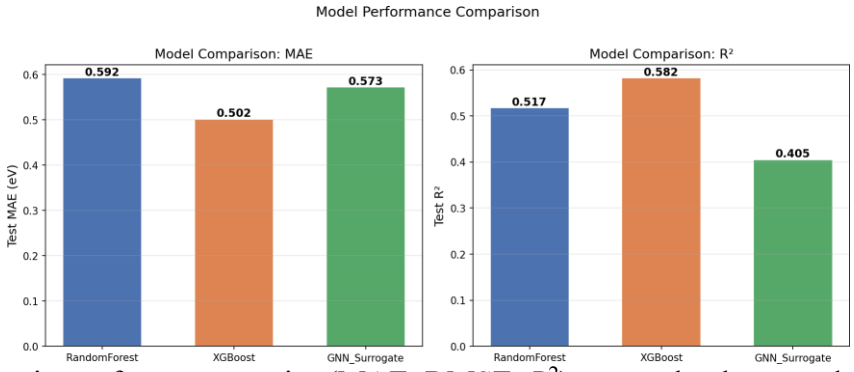


Figure 5: Comparison of test set metrics (MAE, RMSE, R^2) across the three model architectures. XGBoost achieves the lowest MAE and highest R^2.

Character of chemical bonds, which directly influences the electronic band structure and hence the bandgap (Salem and Youssef, 2025; Sharma, 2025).

Table 2: Top 10 features ranked by mean absolute SHAP value for Random Forest and XG-Boost. The composition weighted mean Pauling electronegativity (X_wmean) is the dominant predictor for both models.

Rank	RF Feature	|SHAP|	XGBoost Feature	|SHAP|
1	X_wmean	0.098	X_wmean	0.297
2	mendeleev_no_min	0.070	min_oxidation_wmean	0.140
3	ie1_wmean	0.056	mendeleev_no_min	0.104
4	min_oxidation_wmean	0.053	mendeleev_no_mean	0.087
5	mendeleev_no_wmean	0.049	X_std	0.059
6	group_min	0.045	X_range	0.056
7	mendeleev_no_mean	0.044	volume	0.055
8	X_std	0.038	group_min	0.054
9	X_mean	0.036	lattice_b	0.054
10	mendeleev_no_max	0.036	group_std	0.044

Additional key observations from the SHAP analysis:

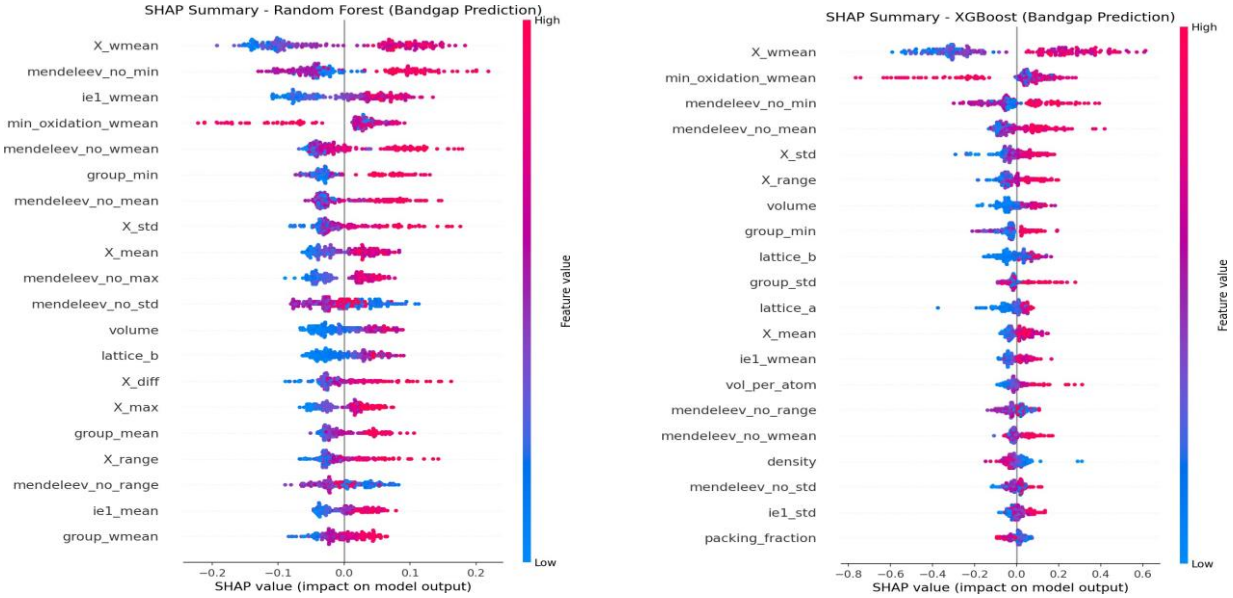
- **Mendeleev number** features (min, mean, max) appear prominently in both models, capturing the periodic table position of constituent elements and correlating with bonding character (ionic vs. covalent).
- **First ionization energy** (ie1_wmean) reflects how tightly valence electrons are bound, di-rectly influencing the valence band position.
- **Electronegativity dispersion** (X_std, X_range, X_diff) captures compositional hetero-geneity materials with more diverse electronegativities tend toward semiconducting behavior.
- **Structural features** (volume, lattice_b) appear in the XGBoost top 10, indicating that lattice geometry contributes to bandgap prediction beyond purely compositional effects.

Figure 6 presents SHAP beeswarm plots for both models. In the XGBoost plot, high values of X_wmean (red dots) consistently push predictions toward higher bandgaps, while low values push toward zero a clear physical signature of the electronegativity bandgap relationship.
Figure 7 shows the corresponding SHAP bar plots (mean absolute SHAP value per feature), providing a complementary global importance view.

Performance Context

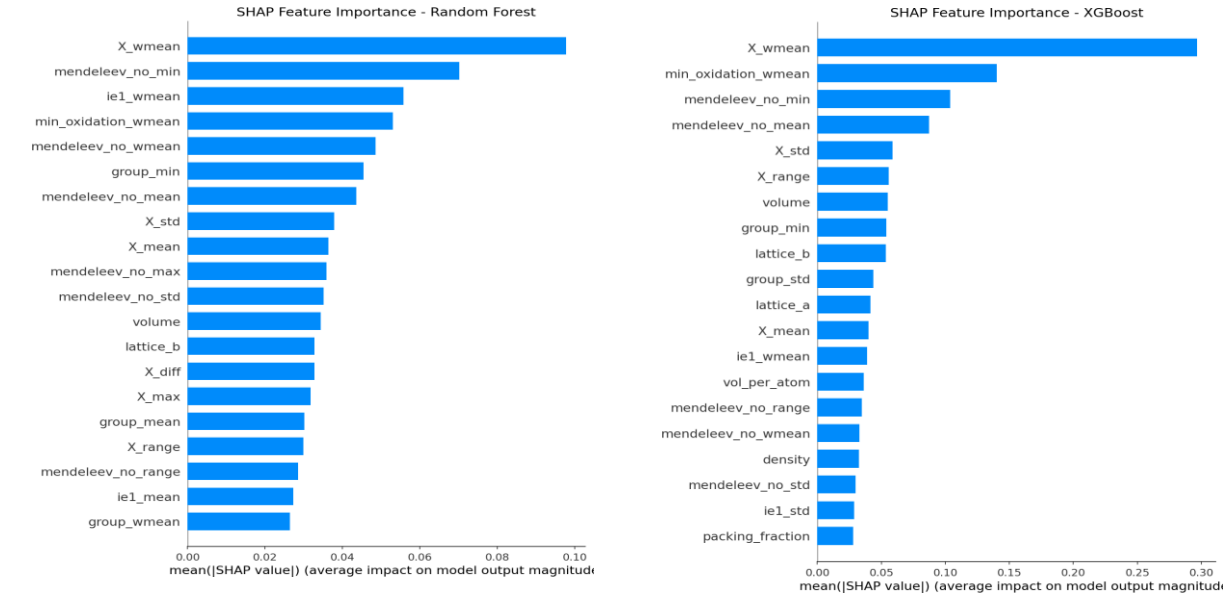
The achieved MAE of 0.502 eV should be interpreted in the context of our deliberately restrictive feature set. For comparison:

- ALIGNN (Choudhary and DeCost, 2021) achieves MAE ≈ 0.20 eV using crystal graph representations on larger JARVIS subsets but requires atomic coordinates and is less interpretable.
- Sharma (Sharma, 2025) reports $R^2 = 0.88–0.90$ on a leakage controlled subset of 2,280 materials, but includes DFT derived descriptors (dielectric tensors) that we intentionally exclude.



(a) Random Forest (b) XGBoost

Figure 6: SHAP beeswarm summary plots for (a) Random Forest and (b) XGBoost. Each dot represents a single test set material; color indicates feature value (red = high, blue = low). The composition weighted mean Pauling electronegativity (X_wmean) is the most influential feature for both models.



(a) Random Forest (b) XGBoost

Figure 7: SHAP bar plots showing mean absolute SHAP values for the top 15 features in (a) Random Forest and (b) XGBoost. Electronegativity related features dominate both models.

- (Zhang et al., 2021) achieve $R^2 > 0.80$ on C2DB with similar compositional features but a larger dataset.

Our lower R^2 (0.582) is therefore attributable to (a) the small dataset size (1,103 materials), (b) the strict exclusion of DFT derived features, and (c) the heavy class imbalance (40% met-als). These are deliberate design choices that prioritize physical interpretability and leakage prevention over raw accuracy.

Virtual Screening and Discussion

Screening Results

Applying the ensemble predictor and scoring function (Eq. 1) to all 1,103 materials, we identify **269 candidates** with predicted bandgaps in the 1.0–2.0 eV optoelectronic window. Table 3 lists the top five candidates ranked by the composite score.

Table 3: Top 5 optoelectronic candidates identified by ensemble based virtual screening. DFT bandgaps are OptB88vdW values from JARVIS-DFT; ensemble predictions are the average of RF, XGBoost, and GNN surrogate models.

Rank	Formula	JVASP ID	E_g^{DFT} (eV)	$\hat{E}_g^{\mathrm{ens}}$ (eV)	Score
1	KTlO	JVASP-8951	1.528	1.504	1.192
2	$CdSbS_2Cl$	JVASP-75178	1.330	1.494	1.187
3	$ZnPS_3$	JVASP-75148	2.182	1.491	1.183
4	BiS_2	JVASP-28270	1.448	1.491	1.182
5	MoS_2Cl_3	JVASP-6967	1.483	1.490	1.180

Candidate Analysis

KTlO (Rank 1). This ternary oxide exhibits the closest DFT ensemble agreement ($\Delta = 0.024$ eV) and a DFT bandgap of 1.528 eV, placing it squarely in the visible light absorption range. The presence of thallium (Tl) suggests strong spin orbit coupling effects that could be exploited for spintronic applications, though toxicity considerations may limit practical deployment.

$CdSbS_2Cl$ (Rank 2). This quaternary chalcohalide has a DFT bandgap of 1.330 eV, slightly below the ensemble prediction of 1.494 eV. The mixed anion chemistry (S^{2-} and Cl^{-}) pro-vides chemical tunability, and related Sb based chalcogenides have shown promise in thin film photovoltaics.

$ZnPS_3$ (Rank 3). This material exhibits the largest DFT ensemble discrepancy ($\Delta = 0.691$ eV), with a DFT bandgap of 2.182 eV but an ensemble prediction of 1.491 eV. This discrepancy warrants further investigation with higher fidelity DFT (HSE06) calculations. $ZnPS_3$ belongs to the metal phosphorus trisulfide family, which has attracted attention for photocatalytic water splitting.

BiS_2 (Rank 4). Bismuth disulfide is a known layered chalcogenide with established applica-tions in thermoelectrics and photodetectors. Its DFT bandgap of 1.448 eV and ensemble pre-diction of 1.491 eV show good agreement, and the material's low toxicity and earth abundant composition make it attractive for sustainable optoelectronics.

MoS_2Cl_3 (Rank 5). This molybdenum based compound is a derivative of the well studied TMDC family. Its DFT bandgap of 1.483 eV is in excellent agreement with the ensemble prediction (1.490 eV). The chlorine functionalization may offer routes to bandgap tuning via halide substitution, analogous to strategies employed in Janus TMDCs.

Discussion

Several observations merit discussion:

1. **Ensemble robustness.** The top five candidates all have ensemble predictions within 0.014 eV of each other (1.490–1.504 eV), suggesting that the scoring function effectively filters for materials near the 1.5 eV optimum. However, the narrow spread also implies that ranking among the top candidates is sensitive to small perturbations, and additional criteria (stability, synthesizability, toxicity) should be incorporated in future work.
2. **DFT ML agreement.** For four of five candidates, the DFT bandgap lies within 0.17 eV of the ensemble prediction. The outlier ($ZnPS_3$, $\Delta = 0.691$ eV) highlights a limitation of com-position only models: materials with unusual bonding environments (e.g., the P–S network in $ZnPS_3$) may not be well captured by statistical aggregates of elemental properties.
3. **Chemical diversity.** The top five span oxides (KTlO), chalcohalides ($CdSbS_2Cl$), phosphorus trisulfides ($ZnPS_3$), binary chalcogenides (BiS_2), and functionalized TMDCs (MoS_2Cl_3), demonstrating that the screening pipeline is not biased toward a single material family.
4. **Limitations.** The moderate R^2 (0.582) means that the ensemble predictions carry substan-tial uncertainty. We recommend that all top candidates be validated with higher fidelity DFT calculations (HSE06 or GW) before experimental synthesis. Additionally, the current pipeline does not assess dynamic stability (phonon spectra) or exfoliation feasibility, which are critical for practical 2D material applications.

Conclusion and Future Work

We have presented an end-to-end, cloud enabled machine learning framework for bandgap prediction and virtual screening of 2D nanomaterials using the JARVIS-DFT dataset. By engi-neering 94 leakage free compositional and structural features and benchmarking three model architectures, we demonstrate that XGBoost achieves the best predictive performance (MAE= 0.502 eV, R^2 = 0.582) among the tested models. SHAP analysis provides physically inter-pretable explanations, identifying the composition weighted mean Pauling electronegativity as the dominant predictor a finding consistent with fundamental chemistry. The ensemble based virtual screening pipeline identifies five promising optoelectronic candidates (KTlO, $CdSbS_2Cl$, $ZnPS_3$, BiS_2, MoS_2Cl_3) with predicted bandgaps near the ideal 1.5 eV target.

Several directions for future work are envisioned:

1. **Crystal graph neural networks.** Replacing the tabular GNN surrogate with a full ALIGNN (Choud-hary and DeCost, 2021) or CGCNN (Xie and Grossman, 2018) implementation that learns directly from atomic coordinates and bond angles is expected to improve R^2 to > 0.85, as demonstrated in the literature.
2. **Multi fidelity learning.** Integrating low fidelity PBE data with high fidelity HSE06 or experimental bandgap measurements through transfer learning or co-kriging could reduce the domain gap between computational predictions and experimental reality (Wang et al., 2026).
3. **Dataset expansion.** Merging the JARVIS-DFT 2D subset with complementary databases (C2DB, Materials Project 2D) would increase the training set size and improve model generalization, as suggested by the non plateaued learning curves (Figure 4).
4. **Active learning.** Implementing Bayesian optimization or uncertainty guided active learning to intelligently select which materials to simulate next via DFT could maximize information gain per computational dollar.
5. **Stability and synthesizability screening.** Augmenting the virtual screening pipeline with phonon stability calculations, exfoliation energy estimates, and synthesizability scores would provide a more holistic assessment of candidate viability.
6. **Class imbalance mitigation.** Stratified sampling, oversampling of semiconductors, or a two stage classifier regressor pipeline (first predicting metal vs. semiconductor, then regress-ing the

bandgap for semiconductors only) could address the 40% metal fraction that biases current predictions.

Data Availability

The JARVIS-DFT dataset is publicly available at https://jarvis.nist.gov/ and via the jarvis-tools Python package (Choudhary et al., 2020). All code, trained models, and pro-cessed data are available at the project repository. The MLflow experiment tracking log and the full top 50 screening candidate list are provided as Supplementary Information.

Acknowledgments

We acknowledge the JARVIS team at NIST for maintaining the open access DFT database, and the developers of scikit learn, XGBoost, SHAP, pymatgen, and MLflow for their open source contributions that made this work possible.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Choudhary, K. and DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185.

Choudhary, K. and Garrity, K. F. (2024). InterMat: accelerating band offset prediction in semiconductor interfaces with DFT and deep learning. *Digital Discovery*, 3:1365–1377.

Choudhary, K., Garrity, K. F., Reid, A. C. E., DeCost, B., Biacchi, A. J., Hight Walker, A. R., Tavazza, F., et al. (2020). The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1):173.

Hey, T., Tansley, S., and Tolle, K. (2009). The fourth paradigm: data-intensive scientific discovery. *Microsoft Research*.

Jia, S. et al. (2024). Deep learning in two-dimensional materials: characterization, prediction, and design. *Frontiers of Physics*, 19(5):53601.

Klimeš, J., Bowler, D. R., and Michaelides, A. (2010). Chemical accuracy for the van der Waals density functional. *Journal of Physics: Condensed Matter*, 22(2):022201.

Lu, B. et al. (2024). When machine learning meets 2D materials: a review. *Advanced Science*, 11(13):2305277.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Manzeli, S., Ovchinnikov, D., Pasquier, D., Yazyev, O. V., and Kis, A. (2017). 2D transition metal dichalcogenides. *Nature Reviews Materials*, 2(8):17033.

Microsoft (2023). Azure Machine Learning documentation. https://learn.microsoft. com/en-us/azure/machine-learning/. Accessed: 2026-05-13.

Nematov, D. and Hojamberdiev, M. (2025). Machine learning–driven materials discovery: a review. *arXiv preprint arXiv:2503.18975*.

Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. V., Grigorieva, I. V., and Firsov, A. A. (2004). Electric field effect in atomically thin carbon films. *Science*, 306(5696):666–669.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., and Ceder, G. (2013). Python materials genomics (pymatgen): A robust, open-source Python library for materials analysis. *Computational Materials Science*, 68:314–319.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,

Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perdew, J. P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865.

Salem, M. and Youssef, M. (2025). Accelerated discovery of 2D optoelec-tronic materials via interpretable graph neural networks. *Preprint*. Avail-able at https://d197for5662m48.cloudfront.net/documents/publicationstatus/286495/preprint_pdf/a015fc81ba1e9dcd7cfd3abdc1b650cf.pdf.

Sharma, G. K. (2025). Leakage-aware bandgap prediction on the JARVIS-DFT dataset: a phase-wise feature analysis. *arXiv preprint arXiv:2512.19732*.

Sibi, H. et al. (2024). Advancing 2D material predictions: superior work function estimation with atomistic line graph neural networks. *RSC Advances*, 14:31768–31778.

Socolof, G. (2024). Band gap prediction of 2D materials with graph neural networks. *Stanford CS224W Blog*. Available at https://medium.com/stanford-cs224w/band-gap-prediction-of-2d-materials-with-graph-neural-networks-f83f9f8dc099.

Wang, H. et al. (2026). Benchmarking bandgap prediction in semiconductors under experimen-tal settings. *arXiv preprint arXiv:2604.25568*.

Wang, Z. and Kenry (2025). Explainable ensemble learning to predict anisotropic nanomaterial band gap. *Materials Chemistry Frontiers*.

Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Ma-terials*, 2(1):16028.

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., et al. (2023). MLflow: An open source platform for the machine learning lifecycle. https://mlflow.org/. Accessed: 2026-05-13.

Zhang, Y. et al. (2021). Bandgap prediction of two-dimensional materials using machine learn-ing. *PLoS ONE*, 16(8):e0255637.