BENCHMARKING BAND GAP PREDICTION FOR SEMICONDUCTOR MATERIALS USING MULTIMODAL AND MULTI-FIDELITY DATA

Haolin Wang[1,2], **Xianyuan Liu**[1,2], **Anna Jungbluth**[3], **Alex Ramadan**[4], **Robert Oliver**[5], **Haiping Lu**[1,2]

[1] Centre for Machine Intelligence, University of Sheffield
[2] School of Computer Science, University of Sheffield
[3] Climate Office, European Space Agency
[4] School of Mathematical and Physical Sciences, University of Sheffield
[5] School of Chemical, Materials and Biological Engineering, University of Sheffield
{haolin.wang, h.lu}@shef.ac.uk

ABSTRACT

The band gap is critical for understanding the electronic properties of materials in semiconductor applications. While density functional theory is commonly used to estimate band gaps, it often underestimates values and remains computationally expensive, limiting its practical usefulness. Machine learning (ML) has become a promising alternative for accurate and efficient band gap predictions. However, existing datasets are limited in data modality, fidelity and sample size, and performance evaluation studies often lack direct comparisons between traditional and advanced ML models. Therefore, a more comprehensive evaluation is needed to make progress towards real-world impacts. In this paper, we developed a benchmarking framework for ML-based band gap prediction to address this gap. We compiled a new multimodal, multi-fidelity dataset from the Materials Project and BandgapDatabase1, consisting of 60,218 low-fidelity computational band gaps and 1,183 high-fidelity experimental band gaps across 10 material categories. We evaluated seven ML models, from traditional methods to graph neural networks, assessing their ability to learn from atomic properties and structural information. To promote real-world applicability, we employed three metrics: mean absolute error, mean relative absolute error, and coefficient of determination R^2. Moreover, we introduced a leave-one-material-out evaluation strategy to better reflect real-world scenarios where new materials have little to no prior training data. Our findings offer valuable insights into model selection and evaluation for band gap prediction across material categories, providing guidance for real-world applications in materials discovery and semiconductor design. The data and code used in this work are available at: https://github.com/Shef-AIRE/bandgap-benchmark.

1 INTRODUCTION

The band gap, defined as the energy difference between the valence and conduction bands, is a fundamental property of periodic solids and plays a critical role in determining their electrical conductivity. This property is widely utilized in semiconductor applications (Yoder, 1996), including light-emitting diodes (LEDs) (Lisensky et al., 1992), transistors (Ueno et al., 2004), and photovoltaic devices (Goetzberger & Hebling, 2000). However, accurately determining the band gap of a material remains a significant challenge. Theoretical methods, such as density functional theory (DFT), are commonly used but often underestimate band gaps due to limitations in exchange-correlation functionals. More advanced methods, such as the G_0W_0 approximation and hybrid functionals (Heyd et al., 2003), provide improved accuracy but are computationally intensive and require meticulous parameter tuning. Recently, machine learning (ML) has emerged as a promising alternative

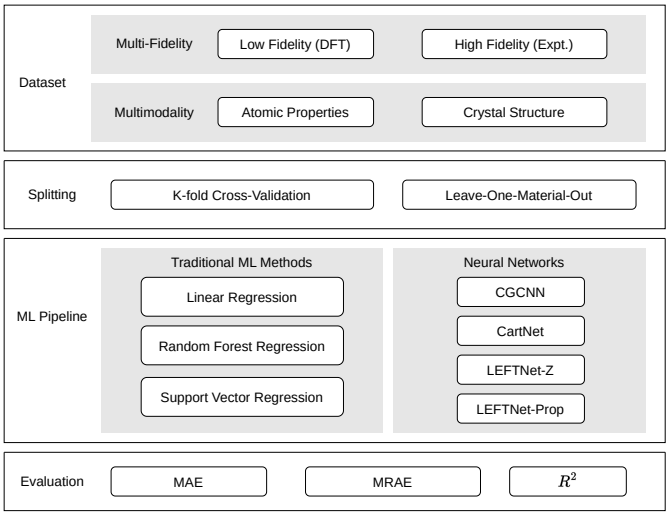


Figure 1: Flowchart of our proposed benchmark. The benchmark categorizes data based on fidelity and modality, incorporating low-fidelity (DFT) and high-fidelity (experimental) band gaps, along with multimodal features. Beyond traditional K-fold cross-validation, a leave-one-material-out strategy is introduced to better reflect real-world scenarios. The machine learning (ML) pipeline studies both traditional ML methods and more recent neural networks. For evaluation, mean relative absolute error (MRAE) is introduced to enhance applicability, alongside mean absolute error (MAE) and the coefficient of determination R^2.

for predicting band gaps. Unlike conventional theoretical methods, ML methods can capture complex structure-property relationships from large datasets, enabling accurate and efficient predictions without expensive calculations.

Machine learning predicts band gaps primarily using two complementary types of information: atomic properties and crystal structure. These modalities represent the material from different perspectives, forming a multimodal data representation (Liu et al., 2025). Atomic properties capture intrinsic characteristics of individual atoms that influence electronic behavior and have been widely used in band gap modeling (Talapatra et al., 2023). For instance, Sabagh Moeini et al. (2024) used eight atomic features to train linear models and identified the standard deviation of valence electrons as a key predictor for band gaps in perovskites.

Advanced graph representation learning models (Schütt et al., 2017; Choudhary & DeCost, 2021) extract crystal structure information via graph neural networks (GNNs) and capture atomic interactions by analyzing distance and orientation. These models better utilize the underlying physics of crystal structures via the three-dimensional arrangement of atoms, making them an intuitive and suitable approach for accurate property prediction. Additionally, structural information complements atomic properties by providing a global context that connects the local information carried by individual atoms. The Crystal Graph Convolutional Neural Network (CGCNN) (Xie & Grossman, 2018) is one of the most widely used models for structure-based materials property prediction. It incorporates nine atomic properties along with interatomic distance information. Subsequent models, such as CartNet (Solé et al., 2025), extend this idea by explicitly encoding the full 3D structure of materials. Another approach, LEFTNet (Du et al., 2023), further improves predictive performance by capturing higher-order geometric features, including bond angles and local orientations. However, none of these methods have been evaluated against traditional machine learning models within a unified benchmark.

Several benchmark studies have compared machine learning models for predicting various material properties. For example, MatBench (Dunn et al., 2020) provides a leaderboard for structure-based property predictions in inorganic materials, covering 13 supervised learning tasks (including band gap prediction) and incorporating both DFT and experimental data. Similar ML benchmarking efforts for band gap prediction include MatDeepLearn (Fung et al., 2021), Varivoda et al. (2023), and the JARVIS-Leaderboard (Choudhary et al., 2024). These benchmarks primarily rely on band

gap databases, the Materials Project (Jain et al., 2013) and QMOF (Rosen et al., 2021; 2022), that provide DFT-calculated band gaps only. On the other hand, experimental datasets, such as the one from Zhuo et al. (2018), contain only compositional information, making them unsuitable for structure-based approaches.

Masood et al. (2023) introduced a multi-fidelity open-access dataset that includes 3D structures, computational band gaps, and experimental band gaps, offering a more suitable resource for structure-based band gap prediction. However, the evaluation dataset is relatively small, containing only 30 materials, and lacks representation of key material categories such as oxides and halides. This limited diversity does not adequately reflect the variety of real-world semiconductor materials, highlighting the need for a more comprehensive dataset that covers a broader range of material classes.

To address these limitations, we introduced a new benchmark that includes a large-scale, multi-fidelity, and multimodal dataset, along with a systematic evaluation of various machine learning methods. The data and code used in this study are available at `https://github.com/Shef-AIRE/bandgap-benchmark` for reproducibility.

Our work makes the following three key contributions:

- We compiled a multi-fidelity dataset from the Materials Project (Jain et al., 2013) and Bandgap-Database1 (Dong & Cole, 2022), comprising 60,218 Perdew–Burke–Ernzerhof (PBE) band gaps and 1,183 experimentally measured band gaps from scientific literature. In this dataset, experimental data were aligned with corresponding 3D structures, enabling the prediction of experimental band gaps using multimodal inputs.
- We compared seven different ML prediction approaches, including traditional machine learning methods and neural networks, in a multimodal (bimodal) setting that incorporates atomic properties and 3D structural information. Additionally, inspired by the pipeline proposed by Masood et al. (2023), we evaluate the effectiveness of multi-fidelity data in predicting experimental band gaps.
- We evaluated each model using three metrics, which are mean absolute error (MAE), mean relative absolute error (MRAE), and coefficient of determination (R^2). Moreover, we introduced a leave-one-material-out strategy to test generalization to unseen material classes. This approach realistically simulates scenarios in which the model encounters entirely new material families, thereby providing a robust assessment of performance and generalization.

2 METHOD

In this study, we designed a benchmark for comparing various ML methods in predicting the band gap of semiconductors. As illustrated in Fig. 1, we considered four key aspects: dataset, data splitting, ML pipeline and evaluation.

2.1 DATASET

High-fidelity experimental data can significantly improve ML model performance, but such data are costly to acquire and have limited availability. In contrast, computational data is more accessible but prone to errors. Masood et al. (2023) proposed a transfer learning framework in which models are first pre-trained on large amounts of low-fidelity computational data to capture general patterns and then fine-tuned using a smaller set of high-fidelity experimental data. This strategy improves predictive performance even when experimental data is scarce.

However, their approach has certain limitations. The evaluation set is relatively small and contains structural overlaps with the training dataset, which may affect the effectiveness of the evaluation.

Building upon their work, we expanded the pre-training dataset by a factor of three and the fine-tuning dataset by a factor of two, ensuring no overlap between the two subsets. Our dataset consists of two parts: computational band gaps extracted from the Materials Project, and experimental band gaps derived from Dong & Cole (2022), which were collected from the literature. Given the larger scale of our fine-tuning set, we expected the transfer learning approach to yield competitive performance. While more advanced domain adaptation techniques (Li et al., 2024) could be explored in

future work, our current setup should provide a practical and scalable baseline that balances simplicity and effectiveness.

Our data collection and curation process has two steps:

PBE Data Filtering. We first collected computational data from the Materials Project, which provides both 3D structures and PBE band gap values. To ensure relevance to semiconductor behaviour, we excluded entries with formulas containing more than eight elements or with band gap values outside the range of 0.5–5 eV. After this filtering process, 61,570 entries remained.

Experimental Data Integration. Next, we sourced experimental band gaps from BandgapDatabase1 (Dong & Cole, 2022), removing records without a DOI to maintain traceability. We matched 39,300 records from BandgapDatabase1 to the Materials Project database based on material formulas. Any entries reporting a band gap range (e.g., 3.0–3.2 eV) were excluded. For materials with multiple entries of the same formula, we took the median band gap value within the 0.5–5 eV range. In cases of isomers, we selected the one with the lowest formation energy, assuming it is the most likely to form. This data filtering method ensures that values largely agree across sources, focusing less on environmental conditions and more on intrinsic material properties.

After processing, 1,183 materials with experimental band gaps were included in the fine-tuning dataset, while 60,218 materials were allocated to the pre-training dataset. Although PBE band gap values exist for the fine-tuning set, we did not use them to avoid assigning different targets to the same input. There is no overlap between these two datasets.

2.2 Machine Learning Models

We adapted the multi-fidelity pipeline from Masood et al. (2023) to leverage both computational and experimental datasets. Specifically, GNN-based models are first pre-trained on the computational (PBE-level) dataset, which provides a large volume of low-fidelity data. This step is important because GNNs typically require substantial amounts of data to learn meaningful representations. The pre-trained models were then fine-tuned on the experimental dataset, which contains high-fidelity band gap measurements.

For comparison, we also trained each GNN-based model from scratch using only the experimental dataset. In the case of traditional ML models, pre-training is not applied, as these simpler architectures rely on feature engineering rather than hierarchical feature learning. Unlike deep neural networks, they do not benefit from large-scale parameter initialization, making pre-training unnecessary (Pan & Yang, 2010).

Our benchmark study evaluated seven machine learning methods, three traditional ML models and four GNN-based models.

Traditional ML Models. We considered three traditional methods: Linear Regression (LR), Support Vector Regression (SVR), and Random Forest Regression (RFR). These methods rely on atomic properties and interatomic distance encoding derived from the raw input data.

Graph Neural Networks. We considered four GNN-based methods: The Crystal Graph Convolutional Neural Network (CGCNN) (Xie & Grossman, 2018), CartNet(Solé et al., 2025) and two variants of LEFTNet(Du et al., 2023). CGCNN represents materials as graphs, where nodes correspond to atoms and edges encode pairwise interatomic distances. This model utilizes structural information solely from these distances, without an explicit representation of the complete three-dimensional spatial relationships. A summary of the key architectural configurations used for these GNN-based methods is provided in Table A5.

CartNet employs Cartesian coordinate-based encoding of crystal structures, capturing higher-order geometric information through absolute positional features. LEFTNet also captures higher-order geometric information, but does so by emphasizing local substructural relationships. It models detailed 3D spatial features such as bond angles and relative atomic arrangements, making it particularly effective for representing localized atomic environments. These enriched representations are valuable for predicting properties influenced by complex three-dimensional interactions.

LEFTNet-Z (the original LEFTNet), which encodes atoms using only their atomic numbers while also integrating 3D structural information, and LEFTNet-Prop where we modify LEFTNet-Z by adopting a more informative one-hot encoding scheme inspired by CGCNN, allowing the network to capture richer atomic-level information in addition to the structural features. By leveraging both explicit atomic properties and structural relationships, LEFTNet-Prop aims to enhance the model's ability to generalize across different material compositions.

2.3 DATA ENCODING

In our benchmark, different models rely on distinct feature representations. In this section, we describe the encoding strategies used for atomic properties and interatomic distances, which serve as the input representations for these models.

Atomic-Level Feature Encoding. To represent atomic properties, we adopted a one-hot encoding scheme over nine features: group number, period number, electronegativity, covalent radius, number of valence electrons, first ionization energy, electron affinity, block, and atomic volume. This feature set—motivated by the design in CGCNN—is applied to all models except for CartNet and LEFTNet-Z. By transforming these atomic properties into categorical variables, our models can effectively discern elemental differences crucial for predicting material properties.

Structural Encoding with Radial Basis Functions (RBF). Structural interactions between atoms are encoded via a set of radial basis functions (RBFs) that transform Euclidean distances into informative edge features. For any pair of atoms i and j with Cartesian coordinates $\mathbf{x}_i$ and $\mathbf{x}_j$, we define the edge attribute as:

$$r_{ij} = \mathbf{RBF}\Big(\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)\Big), \tag{1}$$

where $\mathbf{RBF}(\cdot)$ denotes the collection of radial basis functions. In the CGCNN model, this distance encoding is integrated by concatenating the feature vector of atom i ($\mathbf{v}_i$) with that of its neighbor j ($\mathbf{v}_j$) and the RBF-transformed distance r_{ij}, forming the combined neighbor feature:

$$\mathbf{h}_{ij} = \mathrm{concat}\big(\mathbf{v}_i,\, \mathbf{v}_j,\, r_{ij}\big). \tag{2}$$

These concatenated features are then processed through successive graph convolution layers to yield an atom-level encoding that encapsulates the local chemical environment.

Model-Specific Aggregation. Different architectures use the encoded features in tailored ways:

- **Traditional ML models**: For LR, SVR, and RFR, we adopt a similar encoding approach to CGCNN but concatenate features at the compound level instead of the atom level. A single compound-level representation is obtained by averaging the atom-level encodings $\mathbf{h}_i$ across the entire crystal. This global aggregation provides a fixed-length descriptor for each material.
- **CGCNN**: In CGCNN, we use the original encoding scheme, where concatenated neighbor features are propagated through the network to form atom-level representations.
- **CartNet**: CartNet uses the atomic number as the atomic-level encoding and treats the RBF-encoded distance as a scalar edge attribute. It focuses on atomic positions and addresses rotational equivariance through data augmentation techniques based on Cartesian coordinates.
- **LEFTNet-Z** and **LEFTNet-Prop**: Both models retain the same structural encoding as in the original work, differing only in their atomic-level encoding, as previously described. They treat the RBF-encoded distance r_{ij} as a scalar edge attribute during the message-passing process, preserving local geometric nuances and enabling the networks to learn the interplay between atomic features and interatomic distances.

2.4 DATA SPLITTING

We performed evaluation using two different splitting strategies: cross-validation and leave-one-material-out.

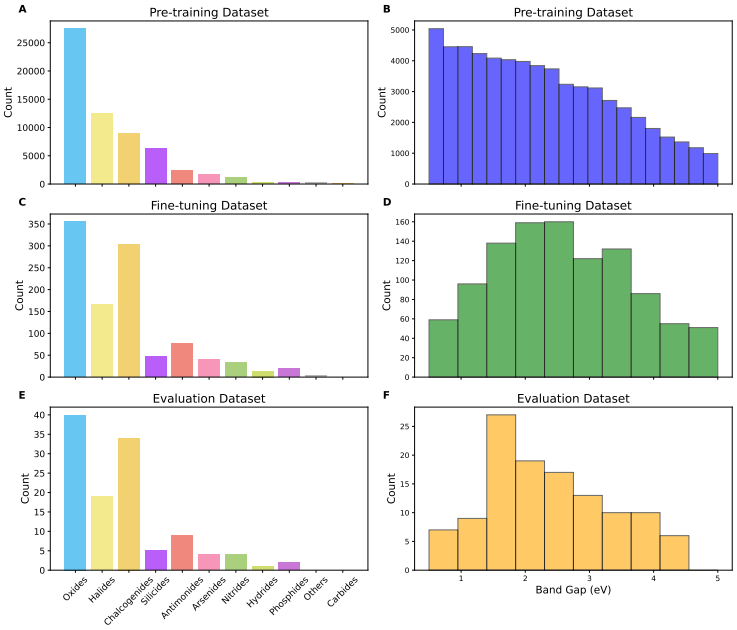


Figure 2: Distribution of datasets used in this study. (A, C, E) show the distribution of material categories in the pre-training, fine-tuning, and evaluation datasets, respectively. (B, D, F) illustrate the corresponding band gap distributions.

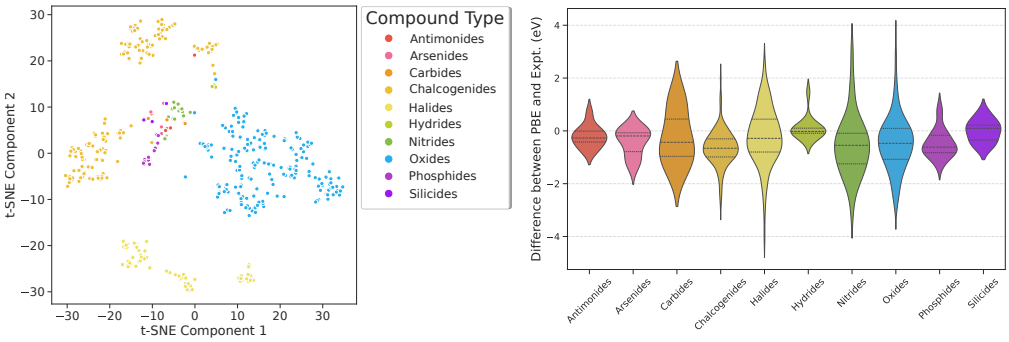
K-fold Cross-Validation Split. The experimental band gap was first split into a fine-tuning set (used for cross-validation) and an evaluation set in a 0.9:0.1 ratio. To ensure both sets are balanced across material categories and maintain similar distributions, we used stratified splitting, randomly selecting data points from each category according to this ratio. Specifically, 118 out of the 1,183 materials were placed in the evaluation set, reserved exclusively for performance testing and not used in any training or validation. The remaining 1,065 materials were used for cross-validation.

Figure 2 illustrates the distribution of material classes and band gaps for each split: panels (A) and (B) correspond to the pre-training dataset, (C) and (D) to the fine-tuning dataset, and (E) and (F) to the evaluation dataset. The distribution between the pre-training dataset and the fine-tuning dataset shows slight differences. Chalcogenides appear more frequently in the experimental datasets, suggesting that band gap measurements for these materials are more commonly available. Additionally, in contrast to the pre-training dataset, which is skewed towards low band gap values, the fine-tuning dataset shows a more centered distribution around 1.5–3 eV.

Leave-One-Material-Out Split. In this experiment, each material was assigned to exactly one category based on the presence of specific elements in its formula (see Table A2). To ensure a clear separation among categories, we excluded 279 out of 1,183 materials that belong to multiple categories or do not fit into any category to avoid ambiguity and potential "leakage" in the leave-one-material-out setting. Table A3 shows the number of materials associated with different category counts. After removing those that belong to more than one category or none, the remaining dataset contains materials that can be uniquely categorized. Figure A2 illustrates the distribution and variability of experimental band gaps within these categories, highlighting their distinct properties.

Table A4 lists the final number of compounds in each category. To evaluate the model's ability to generalize beyond its training distribution, we adopted a leave-one-material-out evaluation protocol. Following the category definitions from Masood et al. (2023), we excluded the "double anion" category, as these materials may overlap with other categories, potentially affecting the evaluation's effectiveness. The dataset is then divided into 10 distinct categories: chalcogenides, oxides, halides, nitrides, phosphides, arsenides, antimonides, silicides, carbides, and hydrides.

For the GNNs training instance, we used the same pre-trained model for each split. One category was held out from the training process, and the model was fine-tuned using all remaining categories. The



(a) (b)

Figure 3: (a) t-SNE visualization of the learned material embeddings, showing distinct clusters for most categories. (b) Violin plot of the difference between PBE-calculated and experimental band gaps (ΔE) across different compound categories, illustrating the variability in each category.

trained model was then tested on the held-out category to assess how well the learned representations transfer to an unseen class of materials.

This evaluation approach simulates a realistic scenario where the model must predict the properties of a new material family with minimal or no prior examples in the training set. It also serves as a rigorous test of whether the model can generalize, evaluates how effectively it captures fundamental patterns in the data, and demonstrates the transferability of the learned representations.

2.5 Evaluation Metrics

We evaluated the regression tasks using three metrics: mean absolute error (MAE), mean relative absolute error (MRAE), and the coefficient of determination (R^2). MAE, a widely used metric in existing works, measures the average magnitude of errors between predicted and actual values, providing an intuitive interpretation of model performance. R^2 quantifies the proportion of total variation of outcomes the model explains. MRAE normalizes the error, making it meaningful regardless of the band gap's magnitude and allowing for fair comparisons across materials with diverse band gap ranges. In particular, for materials with small band gaps, a small absolute error may correspond to a large relative error, which can significantly affect their properties. Therefore, we emphasized MRAE in our experimental setting and selected the best model for each cross-validation fold based on the lowest MRAE.

3 Experiments

3.1 Experimental setup

For traditional ML methods, we encoded the data before using it as input for training. We implemented LR, RFR, and SVR using Scikit-learn version 1.6.0 (Pedregosa et al., 2011). To ensure a fair comparison, we followed the default settings in Scikit-learn without performing any hyperparameter fine-tuning. For the GNNs, we first pre-trained the model on the computational dataset for 200 epochs at a learning rate of 0.01. During the fine-tuning phase, we performed a 10-fold cross-validation on the experimental dataset. To ensure stable gradient updates and mitigate catastrophic forgetting of the pre-trained weights, the learning rate was reduced to 0.001. For models without pre-training, each network was trained directly on the experimental dataset using the same learning rate (0.001) for comparison. The best model for each fold was selected based on the lowest MRAE on the validation set. To ensure the model reaches an optimal state at the end of training, we set the total number of epochs to 50 for pre-trained models and 100 for those without pre-trained models, allowing sufficient convergence for the best model selection. Each method was evaluated using 10-fold cross-validation. In each fold, the validation set was randomly selected from the fine-tuning dataset, while the evaluation dataset served as the test set to assess the performance of the selected best model. For both validation and test sets, we reported the average and standard deviation of each metric across the 10 folds.

Table 1: Performance of different models on the validation and test sets, reported as mean ± standard deviation (**best in bold**, second best in underline). Higher values indicate better performance (↑), otherwise (↓).

Model	Validation set			Test set		
	MAE(eV) ↓	MRAE ↓	R^2 ↑	MAE(eV) ↓	MRAE ↓	R^2 ↑
No pre-training (fine-tuning only)						
LR	$0.664_{(0.008)}$	$0.460_{(0.128)}$	$0.121_{(0.026)}$	$0.709_{(0.055)}$	$0.423_{(0.093)}$	$0.231_{(0.097)}$
RFR	$\underline{0.533}_{(0.010)}$	$\mathbf{0.326}_{(0.007)}$	$\mathbf{0.443}_{(0.022)}$	$0.602_{(0.051)}$	$\underline{0.337}_{(0.060)}$	$\mathbf{0.440}_{(0.054)}$
SVR	$\mathbf{0.524}_{(0.005)}$	$\underline{0.329}_{(0.004)}$	$\underline{0.425}_{(0.010)}$	$0.616_{(0.050)}$	$0.339_{(0.052)}$	$\underline{0.422}_{(0.059)}$
CGCNN	$0.698_{(0.088)}$	$0.353_{(0.056)}$	$0.213_{(0.177)}$	$0.659_{(0.039)}$	$0.364_{(0.009)}$	$0.092_{(0.092)}$
CartNet	$0.660_{(0.035)}$	$0.362_{(0.019)}$	$0.144_{(0.127)}$	$0.721_{(0.063)}$	$0.353_{(0.048)}$	$0.155_{(0.198)}$
LEFTNet-Z	$0.656_{(0.045)}$	$0.348_{(0.054)}$	$0.365_{(0.031)}$	$0.597_{(0.019)}$	$0.357_{(0.016)}$	$0.337_{(0.038)}$
LEFTNet-Prop	$0.653_{(0.053)}$	$0.341_{(0.053)}$	$0.371_{(0.055)}$	$\mathbf{0.575}_{(0.024)}$	$0.351_{(0.020)}$	$0.379_{(0.044)}$
Computational data pre-training + fine-tuning						
CGCNN	$0.684_{(0.067)}$	$0.362_{(0.058)}$	$0.138_{(0.246)}$	$0.626_{(0.048)}$	$0.354_{(0.015)}$	$0.179_{(0.098)}$
CartNet	$0.670_{(0.061)}$	$0.350_{(0.066)}$	$0.282_{(0.088)}$	$0.643_{(0.020)}$	$0.375_{(0.025)}$	$0.251_{(0.042)}$
LEFTNet-Z	$0.654_{(0.058)}$	$0.341_{(0.061)}$	$0.292_{(0.160)}$	$0.608_{(0.043)}$	$\mathbf{0.335}_{(0.014)}$	$0.300_{(0.105)}$
LEFTNet-Prop	$0.642_{(0.059)}$	$0.348_{(0.064)}$	$0.333_{(0.129)}$	$\underline{0.596}_{(0.014)}$	$0.362_{(0.009)}$	$0.342_{(0.030)}$

LR, RFR, SVR, and four pre-trained GNN models were evaluated using the leave-one-material-out strategy. Training was repeated five times, and the mean values of the metrics were recorded. Compared to the MAE, MRAE offers a more meaningful assessment of materials science applications. By normalizing against the actual band gap values, this approach provides clearer insights into the model's performance for both small-gap and large-gap materials, ensuring a more balanced evaluation across the entire band gap range.

3.2 Cross-Validation Results

Table 1 presents the performance of seven different ML models for band gap prediction under two training strategies: (1) direct fine-tuning on experimental data without prior knowledge from computational data and (2) pre-training on computational data followed by fine-tuning on experimental data. All models were evaluated using 10-fold cross-validation, with MRAE as the primary metric for identifying the best-performing model. The results are reported as the mean and standard deviation. RFR and SVR consistently achieve low MRAE values on both the validation and test sets. Among the pre-trained neural networks, LEFTNet-Z generally outperforms CGCNN and CartNet in terms of MRAE across both validation and test data. Notably, LEFTNet-Z's best test set MRAE (0.335 ± 0.014) is lower than most other methods but remains comparable to the RFR results (0.337 ± 0.060). However, when considering R^2 values, RFR and SVR perform significantly better than the neural networks, indicating that there is still room for improvement in deep learning models. These findings suggest that, in our experiments, simpler models such as RFR and SVR serve as strong baseline choices for band gap regression tasks. Overall, pre-training on a large computational dataset helps reduce MRAE when transitioning to real experimental samples. However, RFR trained solely on experimental data proves to be highly effective, often matching or even surpassing the performance of pre-trained GNNs. This suggests that, despite the advantages of deep learning models, traditional machine learning approaches such as RFR and SVR remain strong contenders, particularly when high-quality experimental data is available.

3.3 Leave-one-material-out Results

The MRAE results for different material categories are shown in Figure 4. The MRAE varies significantly across categories, indicating that some materials are inherently more challenging for all models to predict accurately. We also provide the MAE results in Figure A1 to offer additional insights into the models' performance. Silicides and antimonides are among the most challenging categories for all models. This may be due to the diverse bonding characteristics and structural variations of these materials.

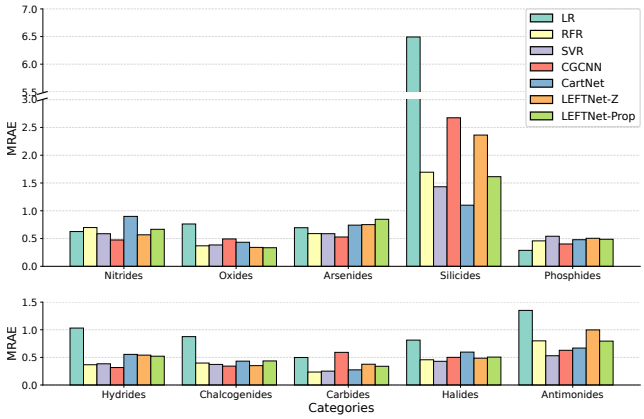


Figure 4: Mean Relative Absolute Error (MRAE) across different material categories in the leave-one-material-out experiments. Each bar represents the performance of a model when a specific material category is excluded from training.

SVR shows consistently leading performance, maintaining moderate errors across most categories, even for those with high MRAE and greater prediction challenges. GNN-based methods generally outperform LR, suggesting that they can better capture features from multimodal data than simpler models. Among them, LEFTNet-Prop achieves the lowest MRAE and MAE, indicating that atomic property encoding and geometric features may be particularly useful for adapting the model to unseen material categories.

4 Conclusion and Discussion

In this work, we presented a benchmarking study for band gap prediction using multi-fidelity and multimodal data, systematically comparing various machine learning approaches from traditional models to graph neural networks (GNNs). Our study addresses the limitations of existing datasets by compiling a comprehensive dataset that integrates 60,218 computational PBE band gaps and 1,183 experimental band gaps, aligned with 3D structural information. This dataset supports structure-based machine learning approaches and provides a stronger foundation for evaluating predictive models in materials science.

Our results highlight that pre-training on computational PBE band gaps can improve the predictive performance of deep learning models on experimental band gaps. However, we also found that traditional models, such as RFR and SVR, remain competitive when trained solely on experimental data. In some cases, these simpler models outperform GNNs, a trend also observed in prior studies such as Dadi et al. (2019). The relatively poor performance of GNNs in our setting may be due to the limited size of the experimental dataset, which restricts their capacity to generalize.

Our leave-one-material-out evaluation revealed that certain material categories, such as silicides and antimonides, pose greater challenges due to their structural and compositional diversity. Additionally, GNNs generally outperform traditional ML models, suggesting their ability to capture complex atomic and structural interactions.

Our evaluation indicates that there is still room for improvement, particularly for neural network models, as their R^2 values remain relatively low. This suggests that the models struggle to capture the variance in the data, potentially limiting their generalization ability. Future work could focus on addressing the limitations identified in the current models to enhance their predictive performance.

Another promising direction is to expand the range of models included in the benchmark. For example, MEGNet, which is designed to handle multi-fidelity data more effectively, could be evaluated to assess whether it better integrates computational and experimental band gaps. Furthermore, expanding the dataset by incorporating additional experimental band gap values from literature and databases could enhance model training and improve generalization across diverse material classes.

REFERENCES

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

Kamal Choudhary, Daniel Wines, Kangming Li, Kevin F. Garrity, Vishu Gupta, Aldo H. Romero, Jaron T. Krogel, Kayahan Saritas, Addis Fuhr, Panchapakesan Ganesh, Paul R. C. Kent, Keqiang Yan, Yuchao Lin, Shuiwang Ji, Ben Blaiszik, Patrick Reiser, Pascal Friederich, Ankit Agrawal, Pratyush Tiwary, Eric Beyerle, Peter Minch, Trevor David Rhone, Ichiro Takeuchi, Robert B. Wexler, Arun Mannodi-Kanakkithodi, Elif Ertekin, Avanish Mishra, Nithin Mathew, Mitchell Wood, Andrew Dale Rohskopf, Jason Hattrick-Simpers, Shih-Han Wang, Luke E. K. Achenie, Hongliang Xin, Maureen Williams, Adam J. Biacchi, and Francesca Tavazza. Jarvis-leaderboard: A large scale benchmark of materials design methods. *npj Computational Materials*, 10(1):93, 2024.

Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, and Gaël Varoquaux. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, 192:115–134, 2019.

Qingyang Dong and Jacqueline M. Cole. Auto-generated database of semiconductor band gaps using ChemDataExtractor. *Scientific Data*, 9(1):193, 2022.

Weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla P Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3d equivariant graph neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G. Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1):84, 2021.

Adolf Goetzberger and Christopher Hebling. Photovoltaic materials, past, present, future. *Solar Energy Materials and Solar Cells*, 62(1–2):1–19, 2000.

Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, 2003.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762, 2024.

George C. Lisensky, R. Lee Penn, Margret J. Geselbracht, and Arthur B. Ellis. Periodic properties in a family of common semiconductors: Experiments with light emitting diodes. *Journal of Chemical Education*, 69(2):151, 1992.

Xianyuan Liu, Jiayang Zhang, Shuo Zhou, Thijs L. van der Plas, Avish Vijayaraghavan, Anastasiia Grishina, Mengdie Zhuang, Daniel Schofield, Christopher Tomlinson, Yuhan Wang, Ruizhe Li, Louisa van Zeeland, Sina Tabakhi, Cyndie Demeocq, Xiang Li, Arunav Das, Orlando Timmerman, Thomas Baldwin-McDonald, Jinge Wu, Peizhen Bai, Zahraa Al Sahili, Omnia Alwazzan, Thao N. Do, Mohammod N. I. Suvon, Angeline Wang, Lucia Cipolina-Kun, Luigi A. Moretti, Lucas Farndale, Nitisha Jain, Natalia Efremova, Yan Ge, Marta Varela, Hak-Keung Lam, Oya Celiktutan, Ben R. Evans, Alejandro Coca-Castro, Honghan Wu, Zahraa S. Abdallah, Chen Chen, Valentin Danchev, Nataliya Tkachenko, Lei Lu, Tingting Zhu, Gregory G. Slabaugh, Roger K. Moore, William K. Cheung, Peter H. Charlton, and Haiping Lu. Towards deployment-centric multimodal AI beyond vision and language. *arXiv preprint arXiv:2504.03603*, 2025.

Hassan Masood, Tharmakulasingam Sirojan, Cui Ying Toe, Priyank V. Kumar, Yousof Haghshenas, Patrick H-L. Sit, Rose Amal, Vidhyasaharan Sethu, and Wey Yang Teoh. Enhancing prediction accuracy of physical band gaps in semiconductor materials. *Cell Reports Physical Science*, 4(9): 101555, 2023.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A.S. Rosen, S.M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J.M. Notestein, and R.Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4:1578–1597, 2021.

A.S. Rosen, V. Fung, P. Huck, C.T. O'Donnell, M.K. Horton, D.G. Truhlar, K.A. Persson, J.M. Notestein, and R.Q. Snurr. High-throughput predictions of metal–organic framework electronic properties: Theoretical challenges, graph neural networks, and data exploration. *npj Computational Materials*, 8:112, 2022.

Alireza Sabagh Moeini, Fatemeh Shariatmadar Tehrani, and Alireza Naeimi-Sadigh. Machine learning-enhanced band gaps prediction for low-symmetry double and layered perovskites. *Scientific Reports*, 14(1):26736, 2024.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017.

Àlex Solé, Albert Mosella-Montoro, Joan Cardona, Silvia Gómez-Coca, Daniel Aravena, Eliseo Ruiz, and Javier Ruiz-Hidalgo. A cartesian encoding graph neural network for crystal structure property prediction: Application to thermal ellipsoid estimation. *Digital Discovery*, 4(3):694–710, 2025.

Anjana Talapatra, Blas Pedro Uberuaga, Christopher Richard Stanek, and Ghanshyam Pilania. Band gap predictions of double perovskite oxides using machine learning. *Communications Materials*, 4(1):46, 2023.

K. Ueno, I. H. Inoue, T. Yamada, H. Akoh, Y. Tokura, and H. Takagi. Field-effect transistor based on KTaO3 perovskite. *Applied Physics Letters*, 84(19):3726–3728, 05 2004.

Daniel Varivoda, Rongzhi Dong, Sadman Sadeed Omee, and Jianjun Hu. Materials property prediction with uncertainty quantification: A benchmark study. *Applied Physics Reviews*, 10(2):021409, 05 2023.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120:145301, 2018.

M.N. Yoder. Wide bandgap semiconductor materials and devices. *IEEE Transactions on Electron Devices*, 43(10):1633–1636, 1996.

Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, 2018.

A Appendix 1: Supplementary Figures

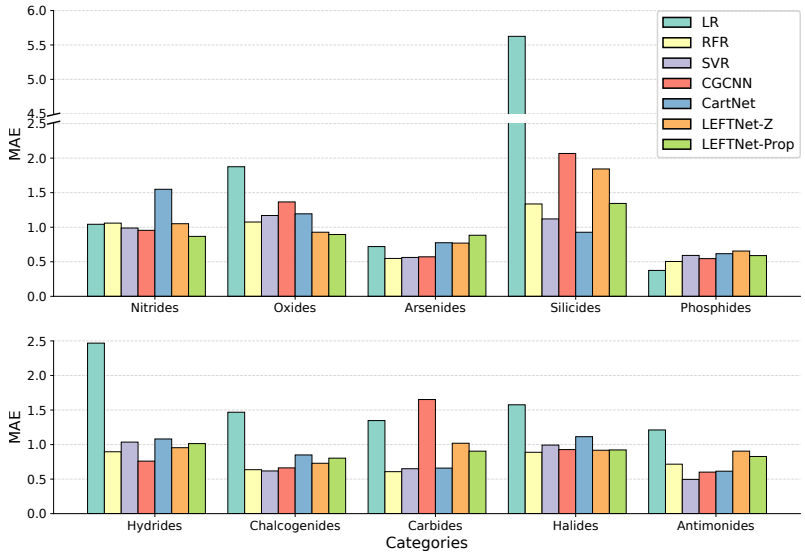


Figure A1: Mean Absolute Error (MAE) across different material categories in the leave-one-material-out experiments. Each bar represents the performance of a model when a specific material category is excluded from training.

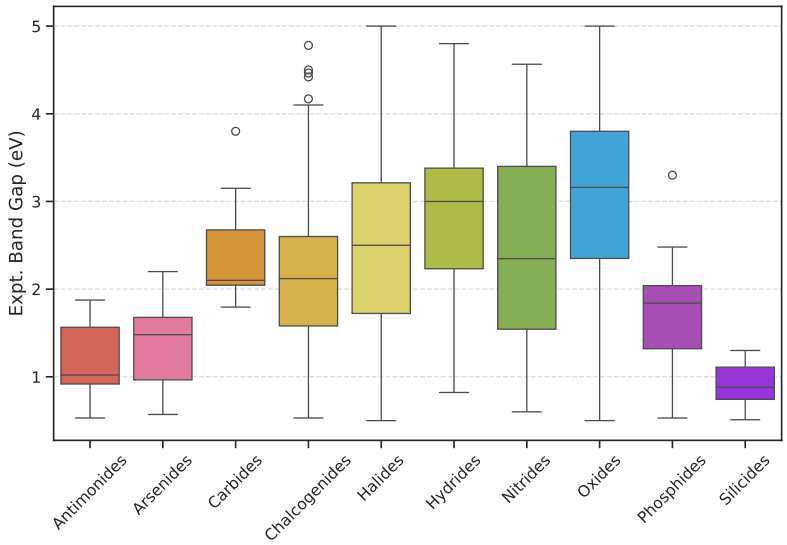


Figure A2: Experimental band gaps across different compound categories, illustrating the distribution and variability within each group.

B Appendix 2: Supplementary Tables

Table A1: Length of one-hot encoded vectors for the elemental properties used in the model.

Atomic property	**Encoding length**
Group	19
Period	7
Electronegativity	10
Covalent radius	10
Valence electrons	12
First ionization energy	10
Electron affinity	10
Block	4
Atomic volume	10

Table A2: Categorization of materials based on chemical composition.

Category	Arsenides	Antimonides	Silicides	Halides	Chalcogenides	Oxides
Key element(s)	As	Sb	Si	F, Cl, Br, I	S, Se, Te	O
Category	Nitrides	Phosphides	Carbides	Hydrides	Others	
Key element(s)	N	P	C	H	Not classified above	

Table A3: Category distribution showing the number of categories each compound belongs to, including those that do not belong to any category.

Category #	1	2	3	Not belong to any
Count	904	263	12	4

Table A4: Number of compounds in each category.

Category	Oxides	Chalcogenides	Halides	Nitrides	Phosphides
Count	387	283	128	34	21
Category	Hydrides	Arsenides	Antimonides	Carbides	Silicides
Count	15	12	10	7	7

Table A5: Key configurations for CGCNN, CartNet, and LEFTNet-Z/Prop.

Parameter	**CGCNN**	**CartNet**	**LEFTNet-Z/Prop**
Batch size	64	32	64
Cutoff radius	8.0 Å	8.0 Å	8.0 Å
Maximum neighbors	12	12	12
RBF dimension	41	64	32
Atom feature dimension	64	256	128
Embedding dimension	128	256	128
Number of layers	3	4	4