

Machine learning prediction of HSE06-level band gaps in two-dimensional semiconductors with reference-guided graph neural networks

Zhen Wan(万振), Shun-Bo Jiang(姜顺波), Yuan Li(李圆), Hui Wang(王辉), Zong-Liang Li(李宗良), and Guang-Ping Zhang(张广平)

Citation: Chin. Phys. B, 2026, 35 (**3**): 037102. DOI: 10.1088/1674-1056/adfef8
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Review of machine learning tight-binding models: Route to accurate and scalable electronic simulations

Jijie Zou(邹暨捷), Zhanghao Zhouyin(周寅张皓), Shishir Kumar Pandey, and Qiangqiang Gu(顾强强)
Chin. Phys. B, 2026, 35 (**1**): 017101. DOI: 10.1088/1674-1056/ae15ef

Classifying extended, localized and critical states in quasiperiodic lattices via unsupervised learning

Bohan Zheng(郑博涵), Siyu Zhu(朱思宇), Xingping Zhou(周兴平), and Tong Liu(刘通)
Chin. Phys. B, 2025, 34 (**1**): 017103. DOI: 10.1088/1674-1056/ad8cb9

Prediction of impurity spectrum function by deep learning algorithm

Ting Liu(刘婷), Rong-Sheng Han(韩榕生), and Liang Chen(陈亮)
Chin. Phys. B, 2024, 33 (**5**): 057102. DOI: 10.1088/1674-1056/ad3908

Quantitative structure-plasticity relationship in metallic glass: A machine learning study

Yicheng Wu(吴义成), Bin Xu(徐斌), Yitao Sun(孙奕韬), and Pengfei Guan(管鹏飞)
Chin. Phys. B, 2021, 30 (**5**): 057103. DOI: 10.1088/1674-1056/abdda5

The effects of combining alloying elements on the elastic properties of γ-Ni in Ni-based superalloy: High-throughput first-principles calculations

Baokun Lu(路宝坤), Chongyu Wang(王崇愚)
Chin. Phys. B, 2018, 27 (**7**): 077104. DOI: 10.1088/1674-1056/27/7/077104

Machine learning prediction of HSE06-level band gaps in two-dimensional semiconductors with reference-guided graph neural networks

Zhen Wan(万振)[1], Shun-Bo Jiang(姜顺波)[1], Yuan Li(李圆)[1], Hui Wang(王辉)[2], Zong-Liang Li(李宗良)[1], and Guang-Ping Zhang(张广平)[1,†]

[1] *Shandong Key Laboratory of Medical Physics and Image Processing & Shandong Provincial Key Laboratory of Light Field Manipulation Physics and Applications, School of Physics and Optoelectronics, Shandong Normal University, Jinan* 250358, *China*
[2] *Shandong Key Laboratory of Space Environment and Exploration Technology & School of Physics and Electronic Engineering, Qilu Normal University, Zhangqiu* 250200, *China*

(Received 9 June 2025; revised manuscript received 19 August 2025; accepted manuscript online 26 August 2025)

Two-dimensional (2D) semiconductors have emerged as promising candidates in next-generation nanoelectronics and sustainable energy technologies, particularly in photoelectrochemical water splitting, due to their exceptional quantum confinement effects and tunable optoelectronic properties. Accurate determination of electronic band gaps remains a critical prerequisite for rational material design in advanced optoelectronic applications. However, the commonly used density functional theory approach with conventional functionals suffers from intrinsic deficiencies in predicting semiconductor band gaps, while calculations with higher hierarchy of functionals like the HSE06 hybrid functional or based on higher level methodologies such as GW approximation incur prohibitive computational costs. To address this challenge, here we propose a reference-guided graph neural network (RG-GNN) framework that achieves HSE06-level accuracy through efficient machine learning. Our approach uniquely embeds an input reference value for the target property with minimal elementary descriptors encoding the structural information of the materials in the model, enabling high-accuracy band gap prediction at the HSE06 level. The model achieves a mean absolute error of 0.15 eV on unseen 2D semiconductor systems compared to HSE06 band gaps. Systematic ablation studies reveal that the reference-guided mechanism reduces prediction error by 83.3% and significantly decreases training dataset requirements for model convergence compared to conventional GNN architectures. Our results demonstrates that topological atomic descriptors from primitive cells, when combined with appropriate reference values, contain sufficient information for highly accurate band gap prediction in 2D materials.

Keywords: machine learning, graph neural network, two-dimensional semiconductors, band gaps, HSE06 accuracy

PACS: 71.15.–m, 07.05.Mh, 02.70.Rr

DOI: 10.1088/1674-1056/adfef8 **CSTR:** 32038.14.CPB.adfef8

1. Introduction

The successful fabrication of graphene by Novoselov, Geim *et al.* in 2004[1] ushered in a new era of materials science, establishing atomically thin two-dimensional (2D) materials as a fundamentally novel class of quantum systems. Since then, 2D materials have rapidly garnered extensive research interest due to their exceptional electronic, optical, and mechanical properties[2–5] and their promising applications in nanoelectronics, optoelectronics, energy storage, biomedicine, and so on.[6,7] In addition to the highly conductive graphene, numerous other 2D semiconductors such as transition metal dichalcogenides (TMDCs), hexagonal boron nitride (h-BN), and phosphorene exhibit unique properties such as tunable band gaps, high thermal conductivity, and superior electron mobility. These characteristics make them ideal candidates for addressing electronic device miniaturization challenges and enabling flexible functional components.[8–15]

The band gap serves as a critical parameter for a semiconductor to determine its suitability and performance for specific electronic and optoelectronic applications.[16–20] Consequently, accurately determining band gaps of 2D semiconductors is essential for advancing their applications. In addition to the experimental direct measurement, theoretical evaluation remains another indispensable way to determine the band gaps of 2D semiconductors, particularly for materials yet to be synthesized in experiments. However, it is now still a formidable challenge to accurately calculate the band gaps of semiconductors. Widely used density functional theory (DFT) approaches employing local density approximation (LDA) or generalized gradient approximation (GGA) functionals suffer from intrinsic deficiencies, systematically underestimating semiconductor and insulator band gaps.[21–24] Although DFT with higher hierarchy of functionals such as HSE06[25–27] or the many-body approaches such as GW method[28] can substantially improve the accuracy of calculated band gaps, their

[†]Corresponding author. E-mail: zhangguangping@sdnu.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

prohibitively high computational costs render them infeasible for high-throughput studies.

In recent years, machine learning (ML) has made rapid and remarkable progress, offering revolutionary solutions to many longstanding scientific challenges.[29–32] For instance, AlphaFold2 by DeepMind innovatively employs an end-to-end neural network to successfully predict the atomic-level three-dimensional (3D) structures of virtually all known proteins ($\sim$ 200 million) solely from amino acid sequences.[29] This breakthrough resolved the "protein folding problem", which has been perplexing the structural biology community for half a century. At the same time, ML approaches, particularly neural network architectures, have demonstrated exceptional capabilities across diverse fields, including analyzing social networks,[33] establishing knowledge graphs,[34] solving partial differential equations,[35] and advancing materials science and functional device design.[36–55]

Significant efforts have also applied ML approaches to semiconductor band gap prediction. For example, in 2018, Xie *et al.* developed a crystal graph convolutional neural network (CGCNN) framework that could learn eight different properties of crystals from the atomic connections.[37] However, its band gap prediction yielded a mean absolute error (MAE) of 0.388 eV against DFT values based on a large training dataset of 16458 materials. In 2020, Zhu *et al.* employed the support vector machine regression (SVR) method to predict the band gaps of a group of phosphorene-isoelectronic 2D semiconductors.[42] Their model achieved high accuracy at the HSE06 level with a root-mean-square error (RMSE) of less than 0.15 eV. Their prediction was based on both PBE-calculated band gaps and elemental properties, including atomic mass, ionization energy, electron affinity, electronegativity, and the electronegativity difference between cations and anions. However, the generalization of this method to other types of 2D semiconductors remains unknown. In 2021, Wan *et al.* proposed an artificial neural network (ANN) model using six computationally efficient descriptors to improve the accuracy of band gaps from DFT with PBE functional (DFT-PBE) to G_0W_0 level.[46] The MAE of predicted band gaps with respect to those from G_0W_0 calculations is 0.35 eV. More recently, Liu *et al.* predicted HSE06 band gaps using ML models based on 309 descriptors, attaining an MAE of 0.142 eV on testing dataset after training on approximately 8300 materials.[47] Zhang *et al.* proposed an ML model with 19 descriptors for HSE06 band gaps of 2D materials, reporting an MAE of 0.30 eV on testing dataset but an MAE of 0.01 eV on training dataset, indicating overfitting of the model.[49] Chen *et al.* developed an ML model to predict the perovskite band gaps based on 136 descriptors, achieving an RMSE of 0.59 eV, with the formation energy descriptor identified as the most critical (reducing RMSE by 10.6%).[51] Liu *et al.* developed an SSE-BGP model with 30 descriptors for HSE06 band gaps of 2D materials by the combination of principal component analysis (PCA) and Autoencoder architectures. The MAE of predicted band gaps with respect to those from HSE06 calculations is 0.262 eV.[53]

Despite these advances, current ML models for band gap prediction exhibit three key limitations: (i) excessive reliance on high-dimensional descriptors, (ii) suboptimal prediction accuracy with elevated MAE/RMSE values, and (iii) demanding large-scale training datasets. Therefore, a descriptor-minimized, data-efficient ML model capable of achieving band gap prediction with high-level accuracy is still highly required, especially for 2D semiconductors of which available datasets are far fewer than bulk crystal data repositories.

To address these challenges, here we present a reference-guided graph neural network (RG-GNN) model that achieves HSE06-level accuracy in predicting band gaps of 2D semiconductors. Our approach requires only the basic structural information of primitive cells combined with a computationally inexpensive DFT-PBE reference band gap value. Crucially, incorporating this PBE reference band gap value significantly enhances prediction accuracy, reducing the MAE against HSE06 values from 0.90 eV to 0.15 eV with an 83.3% accuracy improvement. Furthermore, the RG-GNN model substantially reduces the training dataset size required for model convergence.

2. Computational details

2.1. Data preparation and graph representation

To establish an ML model for predicting band gaps of 2D semiconductors, we retrieved our dataset from the Computational 2D Materials Database (C2DB),[56,57] a widely used repository containing first-principles calculated properties of already known 2D materials. The dataset comprises 3128 unique 2D materials with calculated HSE06 band gaps. As illustrated in Fig. 1, there are 968 metallic materials (with band gaps of 0 eV) and 2160 semiconductors and insulators. We classified the 2160 non-metallic 2D materials into 11 groups according to their band gap values with an interval of 1 eV for each group, resulting in 292 materials in the (0, 1) eV, 598 materials in [1, 2) eV, 567 materials in [2, 3) eV, 297 materials in [3, 4) eV, 176 materials in [4, 5) eV, 140 materials in [5, 6) eV, 54 materials in [6, 7) eV, 21 materials in [7, 8) eV, and 15 materials with band gaps no less than 8 eV. As 2D semiconductors with band gaps greater than or equal to 8 eV are relatively rare in both routine research and the C2DB database ($< 0.5\%$), we excluded these 15 materials from both the training and testing datasets during model construction. However, in the final stage of our study, we used these 15 materials as

an additional testing dataset to validate the transferability of our RG-GNN model. Namely, we only used the 3113 materials, including 968 metals and 2145 semiconductors with band gaps below 8 eV, to constitute the final datasets, which were randomly divided into the training dataset with 2800 data, validation dataset with 113 data, and testing dataset with 200 data unless otherwise mentioned. Given the limited size of the overall dataset, we allocated a larger proportion of data to the training dataset to ensure model convergence during training. Although the testing dataset comprised 6.4%, less than 10%, of the total data, cross-validation was finally performed to verify the reliability of reported performance metrics for our model.

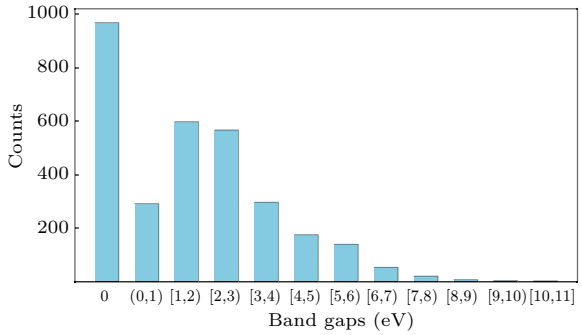


Fig. 1. The distribution of the HSE06 band gaps for the 3128 kinds of 2D materials retrieved from C2DB.

The structural information of each material in the dataset was first transformed into graph data for model input. The input node attributes included the corresponding atomic numbers in the periodic table of elements and the x–y–z Cartesian coordinates of each atom. The edges between nodes were defined using a radial cutoff criterion of 4 Å. All atom pairs within this distance threshold in the primitive cell were connected, establishing edges representing local chemical interactions with neighborhoods for the nodes. To ensure complete edge enumeration under periodic boundary conditions, lattice parameters of the primitive cell were incorporated as global graph attributes. In addition, the global graph attribute can optionally contain a reference value of the band gap.

2.2. Model building

As a widely used type of graph neural networks (GNNs), graph convolutional networks (GCNs)[58] have gained prominence in social network analysis, recommender systems, bioinformatics, and crystal material property prediction owing to their computational efficiency and architectural simplicity. The GCNs also demonstrate advantages in directly processing graph data with irregular structures and effectively capturing hierarchical patterns through simultaneous aggregation of local node neighborhood features and global graph topology information. However, GCNs have the limitation of assigning fixed normalized weights to all neighboring nodes during aggregation, which results in failure of prioritizing nodes based on their significance. This issue can be complementarily solved by graph attention networks (GATs),[59] which leverage attention mechanisms to dynamically assign differentiated attention weights to neighboring nodes and significantly enhance the capability to capture spatial dependencies and structural patterns in graph data. Therefore, here we used the combination of GCN and GAT modules to construct the convolutional layers in our model that are more capable of understanding the structural characteristics, complex atomic interactions, and different significance of neighboring atoms in the primitive cell of 2D materials, which is essential for further exploring the relationship between the atomic structure and band gap for semiconductors.

As illustrated in Fig. 2, the RG-GNN model comprises five core components: input, convolutional layers, pooling layer, full connected layer, and the output. The input consists of nodes representing atoms and edges representing connections between atoms in the primitive cell, which are characterized by vectors corresponding to the atoms and bonds in the 2D materials, respectively. Besides, a reference value of the band gap can be optionally provided as a global graph attribute for the input. One can obtain the predictions from the model after the input graph data undergoing convolutional operations across N hidden convolutional layers followed by pooling and fully-connected operations.

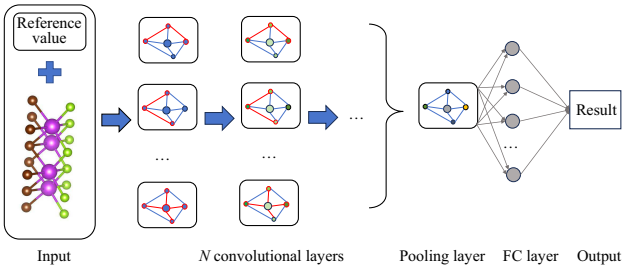


Fig. 2. Schematic diagram of the reference-guided graph neural network (RG-GNN) model. The input consists of the nodes and edges of graph data as well as global graph attributes, which undergoes convolutional operations across N hidden convolutional layers followed by pooling and fully-connected operations to generate predictions.

The depth of convolutional layers in the model critically influences the smoothness, distinctiveness, and gradient of the node features, and ultimately the model prediction accuracy. During the model construction, we systematically evaluated the model with 2 GCN∼5 GCN or GAT convolutional layers. Test results demonstrate optimal model performance at 3-layer depth with 1-layer GCN module and 2-layer GAT modules, which are adopted in subsequent implementations. We further optimized activation functions across seven commonly used candidates: Linear, ReLU, Sigmoid, Tanh, Softplus, ELU, and LeakyReLU, among which ReLU gives the best model prediction performance. Finally, we evaluate the effects of distinct normalization strategies in each convolutional layer on model prediction performance, including graph normalization, batch normalization, layer normalization, and no normalization. It

is showed that the combined application of graph normalization and batch normalization yielded synergistic benefits, accelerating model convergence while reducing prediction error. Therefore, this dual-normalization scheme was implemented across all convolutional layers in our RG-GNN model.

The model prediction accuracy is quantified using the mean absolute error (MAE), error standard deviation (σ), and coefficient of determination (R^2), which respectively defined as

$$\mathrm{MAE} = \overline{\Delta y} = \frac{1}{n}\sum_{i=1}^{n}\Delta y_i = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{1}$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[\Delta y_i - \overline{\Delta y}]^2}, \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3}$$

where y_i is the value of band gap calculated from HSE06 functional, $\bar{y}$ is the mean value of HSE06 band gaps in the sample dataset, $\hat{y}_i$ is the predicted band gap by the RG-GNN model, and n is the size of the sample dataset.

3. Results and discussion

3.1. Model with no reference value for band gap

First, we evaluated the model performance with reference band gap values disabled in global graph attributes. The training dataset size required for model convergence was systematically investigated. As shown in Fig. 3(a), though the model seems converged when the training dataset size reaches 1000, the MAE on testing dataset still shows significant fluctuations, showing an insufficient training dataset of 2800 for robust convergence of the model. On the other hand, the MAE of testing dataset is much larger than that of training dataset, exhibiting overfitting of the model. Furthermore, as shown in Fig. 3(b), the model gets converged slowly during the training, which requires at least of 600 epochs. The MAE for both training and testing datasets are still large, which are 0.58 eV and 0.90 eV, respectively, indicating a poor predictive capability. The poor prediction performance can be more starkly seen from Figure 3(c), where the maximum prediction error reaches 6.03 eV. Therefore, this model fails to achieve high-accuracy predictions.

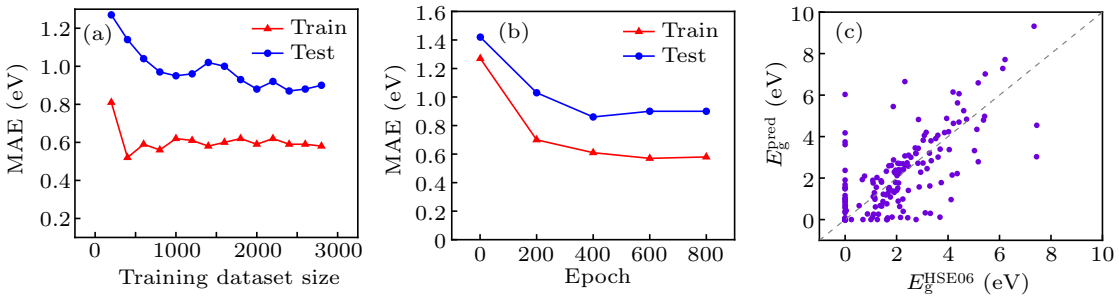


Fig. 3. The performance of constructed model with no reference value for the band gap. (a) The MAE for training and testing datasets as the size of the training dataset increases from 200 to 2800 data by a step of 200 data. (b) The MAE for training dataset and testing dataset as a function of epochs. (c) Parity plot of E_g^{pred} versus E_g^{HSE06} on the testing dataset.

3.2. Model with an interval-related reference value for band gap

Neural network training fundamentally represents a high-dimensional optimization problem, where the core challenge involves navigating complex parameter landscapes. Through iterative parameter updates, such training processes gradually converge toward a global minimum for the loss function. Notably, incorporating domain-specific prior knowledge as reference values can significantly enhance search efficiency in parameter space.[60,61] This approach proves particularly valuable for constructing high-accuracy models under data scarcity constraints.

Inspired by this, we incorporated a reference band gap value into the model input as global graph attribute to improve the prediction accuracy. For this aim, the 2800 data of the training dataset were grouped using three grouping strategies according to their HSE06 band gap values: (i) the training dataset are divided into 2 groups with a group interval of 4.0 eV, the first group has band gaps in [0, 4) eV, and the second one has band gaps in [4, 8) eV; (ii) the training dataset are divided into 4 groups with a group interval of 2.0 eV; and (iii) the training dataset are divided into 8 groups with a group interval of 1.0 eV. During the model training, the lower bound of each data group served as the input reference value for all the data in this group. For example, a reference value of 0 was used for all the data with band gaps in the group of [0, 4) eV while 4.0 eV was used as the reference value for all the data with band gaps in the group of [4, 8) eV.

Figure 4(a) reveals the case where the group interval is 4.0 eV. It is easily seen that the model performance is very similar to the case where the model has no reference value as input for the band gap. The trained model still exhibits deficiencies of overfitting and inadequate predictive capability when the input reference values demonstrate significant deviation from the target HSE06 value. Specifically, the MAE for training and testing datasets are 0.57 eV and 0.76 eV, respectively. However, when the interval of the group is decreased, as shown in Figs. 4(b) and 4(c), the overfitting issue for the

model is resolved. At the same time, the predictive capability is significantly enhanced. The MAE for both training and testing datasets are around 0.37 eV and 0.16 eV for the cases where the group intervals are 2.0 eV and 1.0 eV, respectively. That is to say, for the case with a group interval of 1.0 eV, the RG-GNN model demonstrates exceptionally high prediction accuracy, which currently achieves the highest prediction accuracy among existing descriptor-minimized models.

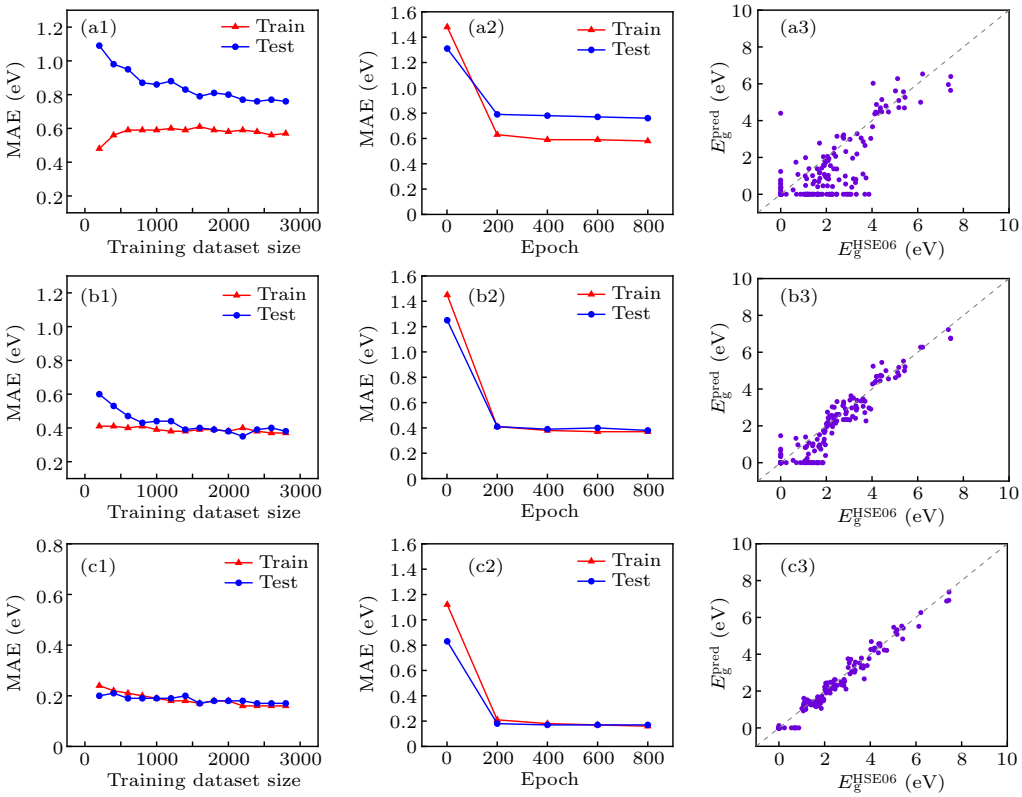


Fig. 4. The performance of RG-GNN model with an interval-related reference value for the band gap, where the 2800 data in the training dataset are categorized into groups with different intervals according to their band gap values. (a) The interval for each group is 4.0 eV, (b) the interval for each group is 2.0 eV, and (c) the interval for each group is 1.0 eV. Panels (a1), (b1), (c1) are MAE for training and testing datasets as the size of the training dataset increases from 200 data to 2800 data by a step of 200 data. Panels (a2), (b2), and (c2) are MAE for training dataset and testing dataset as a function of epochs. Panels (a3), (b3), and (c3) are parity plots of E_g^{pred} versus E_g^{HSE06} on the testing dataset.

3.3. Model with band gap from PBE functional as a reference value

The preceding analysis confirms that appropriate band gap reference values serve as pivotal factors in achieving high prediction accuracy for the model. However, in practical applications especially involving novel materials, prior knowledge of adequate band gap reference values remains unavailable. In such scenarios, we propose utilizing band gaps calculated from the PBE functional as computationally efficient reference inputs. For a target 2D material with PBE band gap already available, it only spends ca. 0.01 s for the well-trained RG-GNN model to predict an HSE06 level band gap. If no PBE band gap is available, the RG-GNN model primarily spends time calculating it. This PBE calculation is typically 1–2 orders of magnitude computationally cheaper than the corresponding HSE06 band gap calculation.[62,63] Therefore, RG-GNN model substantially reduces computational demands compared to HSE06 calculations while maintaining prediction fidelity.

As shown in Fig. 5(a), compared to the case where the model with no reference value for the band gap, when the band gap calculated from PBE functional is incorporated as a reference value, the RG-GNN model gets well converged after training on a training dataset of around 1000. Meanwhile, the MAE on the training dataset and testing dataset are very close and small. This indicates that reference-guided learning successfully captures the fundamental structure-band gap relationship in 2D semiconductors. Furthermore, one can see from Fig. 5(b), after introducing a reference value for the band gap, the model gets converged quickly during the training, which requires only 200 epochs. At the same time, the MAE for both training and testing datasets get smaller, which are around 0.15 eV. The σ and R^2 on the testing dataset are 0.32 eV and 0.96, respectively. The good fitting between E_g^{HSE06} and E_g^{pred} on the testing dataset shown in Fig. 5(c) suggests that this model possesses strong generalization capability for HSE06-level accuracy band gap prediction in 2D materials. Figure 5(d) gives the MAE of E_g^{PBE} and E_g^{pred} relative to

E_g^{HSE06} as a function of band gap. It can be clearly seen that the deviation of PBE band gaps from HSE06 ones increases with the band gap while it does not change much with the material band gap for the RG-GNN model predicted band gaps relative to HSE06 ones. This suggests that the RG-GNN model has a higher relative accuracy for materials with larger band gaps.

Table 1 systematically compares model performance across distinct strategies for the reference values. One can easily see that the reference value for the band gap plays an important role in improving the prediction accuracy of the model. The MAE of the model get smaller as the reference value for the band gap become more appropriate. And the reference values obtained from PBE functional prove to be practically good enough for predicting the HSE06 band gaps.

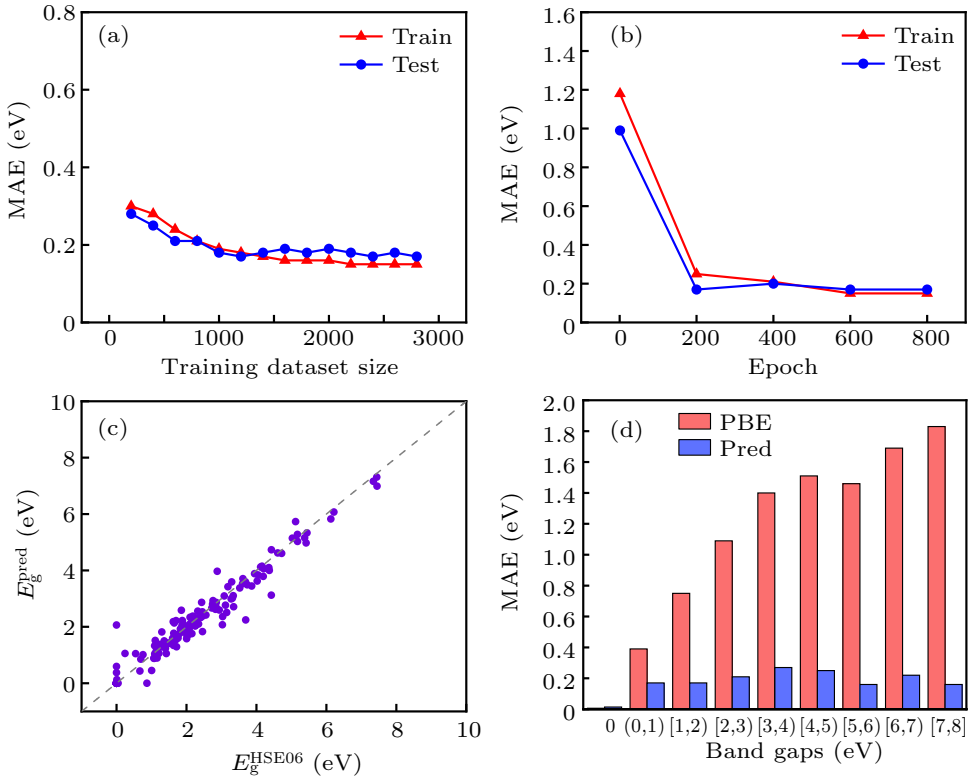


Fig. 5. The performance of RG-GNN model with the band gap from PBE functional as a reference value. (a) The MAE for training and testing datasets as the size of the training dataset increases from 200 data to 2800 data by a step of 200 data. (b) The MAE for training dataset and testing dataset as a function of epochs. (c) Parity plot of E_g^{pred} versus E_g^{HSE06} on the testing dataset. (d) The MAE of E_g^{PBE} and E_g^{pred} relative to E_g^{HSE06} as a function of band gap.

Table 1. The MAE on the testing dataset for the models using different strategies for the band gap reference values.

Strategy for band gap reference value	MAE (eV)
With no reference value	0.90
Interval-based reference value (interval = 4.0 eV)	0.76
Interval-based reference value (interval = 2.0 eV)	0.37
Interval-based reference value (interval = 1.0 eV)	0.16
E_g^{PBE} as the reference value	0.15

To further validate the prediction accuracy and stability of the RG-GNN model, we perform 10-fold cross-validation with the same sizes for the training, validation, and testing datasets as above in each fold. As summarized in Table 2, the MAE values across all 10 folds are consistently around 0.16 eV, with corresponding σ below 0.43 eV and R^2 above 0.91. This consistent performance demonstrates little sensitivity to partitioning randomness as well as the robustness of the prediction accuracy and stability of the RG-GNN model.

During the training and testing datasets construction, 15 materials with band gaps exceeding 8.0 eV were excluded, for which the MAE, σ, and R^2 of E_g^{PBE} versus E_g^{HSE06} are 2.55 eV, 0.24 eV, and −7.63, respectively. To evaluate the transferability and robustness of our RG-GNN model with PBE reference values, we predicted the band gaps E_g^{pred} for these high-band-gap materials.

Table 2. The model performance metrics (MAE, σ, and R^2) in 10-fold cross-validation.

Fold index	MAE (eV)	σ (eV)	R^2
1	0.15	0.32	0.96
2	0.17	0.41	0.93
3	0.15	0.34	0.96
4	0.18	0.43	0.92
5	0.17	0.43	0.93
6	0.15	0.36	0.96
7	0.16	0.42	0.95
8	0.18	0.41	0.91
9	0.15	0.36	0.96
10	0.15	0.33	0.96

Table 3. The predicted band gaps E_g^{pred} for 15 high-band-gap 2D materials with band gaps exceeding 8.0 eV. E_g^{PBE} and E_g^{HSE06} are band gaps from PBE and HSE06 functionals, respectively. The prediction deviation ΔE_g is defined as $\Delta E_g = E_g^{pred} - E_g^{HSE06}$.

Material	E_g^{PBE} (eV)	E_g^{HSE06} (eV)	E_g^{pred} (eV)	ΔE_g (eV)	$\lvert\Delta E_g/E_g^{HSE06}\rvert$ (%)
MgF_2	7.53	9.80	9.70	−0.10	1.0
Y_2F_6	7.73	9.89	9.87	−0.02	0.2
Al_2F_6	8.41	10.79	10.82	0.03	0.2
Ga_2F_6	5.83	8.15	8.10	−0.05	0.6
Sc_2F_6	6.88	10.02	9.01	−1.01	10.0
$HfZnF_6$	5.93	8.57	8.41	−0.16	1.8
MgB_2F_8	8.24	10.62	10.58	−0.04	0.3
SrP_2F_{12}	6.87	9.43	9.44	0.01	0.1
BaP_2F_{12}	7.04	9.58	9.59	0.01	0.1
$Ta_2Na_2F_{12}$	6.24	8.85	8.84	−0.01	0.1
$Ta_2Li_2F_{12}$	6.14	8.73	8.74	0.01	0.1
$Ta_2K_2F_{12}$	6.08	8.68	8.29	−0.39	4.5
$Nb_2Na_2F_{12}$	5.53	8.25	7.84	−0.41	4.9
$Nb_2Li_2F_{12}$	5.44	8.14	7.71	−0.43	5.2
$Nb_2K_2F_{12}$	5.49	8.18	7.73	−0.45	5.5

From Table 3, it is remarkably found that the E_g^{pred} still has a high accuracy compared with E_g^{HSE06} and the MAE of E_g^{pred} versus E_g^{HSE06} dramatically decreases to only 0.21 eV with corresponding σ and R^2 being 0.28 eV and 0.85, respectively. More specifically, the prediction error falls below 5.5% for 14 materials. For Sc_2F_6, the prediction error reaches as large as 1.01 eV (10.0%). Notably, for structurally similar materials including Y_2F_6 (0.2%), Al_2F_6 (0.2%), and Ga_2F_6 (0.6%), the RG-GNN model manifests a high accuracy. We have noted that there is large difference in band structures for Sc_2F_6 compared to these three structural similar counterparts. That is, the d-orbitals make significant contributions to the CBM of Sc_2F_6, indicating strong electron correlation effects in Sc_2F_6. The large deviation for the predicted band gap relative to HSE06 band gap for Sc_2F_6 suggests the RG-GNN model with minimal elementary descriptors is not adequate to accurately predict band gaps of strongly correlated systems. It should be noted that, despite this, the predicted band gap E_g^{pred} (9.01 eV) of Sc_2F_6 is still much closer to E_g^{HSE06} (10.02 eV) than E_g^{PBE} (6.88 eV). Meanwhile, the R^2 for these 15 materials is 0.85, which is smaller than 0.9. Since all these materials have band gaps larger than 8 eV and have been never seen during the model training, a value of 0.85 for R^2 still demonstrates the good prediction accuracy, transferability and robustness of the RG-GNN model with PBE band gaps as reference values. Based on this, it can be naturally speculated that our RG-GNN model can make predictions on a much higher level, such as GW level, as long as that high precision training datasets and appropriate band gap reference values are provided. Hence, our RG-GNN model establishes a new paradigm for high-accuracy-fidelity band gap prediction in 2D semiconductors.

4. Conclusions

In summary, we have proposed an RG-GNN framework that achieves HSE06-level accuracy in predicting band gaps of 2D semiconductors. By embedding computationally efficient PBE band gap references based on minimized descriptors of only fundamental structural information, the model attains an exceptional MAE of 0.15 eV for predicted band gaps against HSE06 values. The reference-guided mechanism significantly reduces the prediction errors by 83.3% and the size of training dataset for model convergence. Our findings demonstrate that topological atomic descriptors preserve sufficient physical determinism for high-fidelity band gap prediction when coupled with appropriate reference values. This work establishes a new paradigm where target-property references fundamentally enhance prediction accuracy of the ML models while streamlining their complexity across computational materials science.

Acknowledgement

Project supported by the National Natural Science Foundation of China (Grant No. 22173052).

References

[1] Novoselov K S, Geim A K, Morozov S V, Jiang D E, Zhang Y S, Dubonos S V, Grigorieva I V and Firsov A A 2004 *Science* **306** 666

[2] Geim A K and Novoselov K S 2007 *Nat. Mater.* **6** 183

[3] Castro Neto A H, Guinea F, Peres N M R, Novoselov K S and Geim A K 2009 *Rev. Mod. Phys.* **81** 109

[4] Splendiani A, Sun L, Zhang Y B, Li T S, Kim J, Chim C Y, Galli G and Wang F 2010 *Nano Lett.* **10** 1271

[5] Mak K F, Lee C G, Hone J, Shan J and Heinz T F 2010 *Phys. Rev. Lett.* **105** 136805

[6] Geim A K 2009 *Science* **324** 1530

[7] Wang Q H, Kalantar-Zadeh K, Kis A, Coleman J N and Strano M S 2012 *Nat. Nanotechnol.* **7** 699

[8] Dean C R, Young A F, Meric I, Lee C, Wang L, Sorgenfrei S, Watanabe K, Taniguchi T, Kim P, Shepard K L and Hone J 2010 *Nat. Nanotechnol.* **5** 722

[9] Bertolazzi S, Brivio J and Kis A 2011 *ACS Nano* **5** 9703

[10] Radisavljevic B, Radenovic A, Brivio J, Giacometti V and Kis A 2011 *Nat. Nanotechnol.* **6** 147

[11] Butler S Z, Hollen S M, Cao L Y, Cui Y, Gupta J A, Gutiérrez H R, Heinz T F, Hong S S, Huang J X, Ismach A F, Johnston Halperin E, Kuno M, Plashnitsa V V, Robinson R D, Ruoff R S, Salahuddin S, Shan J, Shi L, Spencer M G, Terrones M, Windl W and Goldberger J E 2013 *ACS Nano* **7** 2898

[12] Chhowalla M, Shin H S, Eda G, Li L J, Loh K P and Zhang H 2013 *Nat. Chem.* **5** 263

[13] Kou L Z, Chen C F and Smith S C 2015 *J. Phys. Chem. Lett.* **6** 2794

[14] Li X and Zhu H W 2015 *J. Mater.* **1** 33

[15] Ling X, Wang H, Huang S X, Xia F N and Dresselhaus M S 2015 *Proc. Natl. Acad. Sci. USA* **112** 4523

[16] Chaves A, Azadani J G, Alsalman H, Da Costa D, Frisenda R, Chaves A, Song S H, Kim Y D, He D W, Zhou J D, Castellanos Gomez A, Peeters F M, Liu Z, Hinkle C L, Oh S H, Ye P D, Koester S J, Lee Y H, Avouris P, Wang X R and Low T 2020 *npj 2D Mater. Appl.* **4** 29

[17] Weng M Y, Pan F and Wang L W 2020 *npj Comput. Mater.* **6** 33

[18] Oh Y, Song S and Bae J 2024 *Int. J. Mol. Sci.* **25** 13104

[19] Bao X T, Shi J W, Han X, Wu K M, Zeng X, Xia Y X, Zhao J H, Zhang Z Y, Du W N, Yue S, Wu X X, Wu B, Huang Y, Zhang W K and Liu X F 2025 *Nano Lett.* **25** 2639

[20] Madani M, Lacivita V, Shin Y and Tarakanova A 2025 *npj Comput. Mater.* **11** 15

[21] Sadeghi S N, Zebarjadi M and Esfarjani K 2019 *J. Mater. Chem. C* **7** 7308

[22] Hu X M, Yang J L, Wang W, Zhang X J, Meng Y F, Ye Y F, Ding K N, Zhang F J and Zhang S L 2024 *J. Mater. Chem. C* **12** 15662

[23] Alaal N and Roqan I S 2018 *ACS Appl. Nano Mater.* **2** 202

[24] Liu T Y, Zhao X, Liu X F, Xiao W J, Luo Z J, Wang W T, Zhang Y F and Liu J C 2023 *J. Energy Chem.* **81** 93

[25] Heyd J, Scuseria G E and Ernzerhof M 2003 *J. Chem. Phys.* **118** 8207

[26] Heyd J, Scuseria G E and Ernzerhof M 2006 *J. Chem. Phys.* **124** 219906

[27] Krukau A V, Vydrov O A, Izmaylov A F and Scuseria G E 2006 *J. Chem. Phys.* **125** 224106

[28] Shishkin M and Kresse G 2007 *Phys. Rev. B* **75** 235102

[29] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S A A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior A W, Kavukcuoglu K, Kohli P and Hassabis D 2021 *Nature* **596** 583

[30] Samatov M R, Liu D, Zhao L, Kazakova E A, Abrameshin D A, Das A, Vasenko A S and Prezhdo O V 2024 *J. Phys. Chem. Lett.* **15** 12362

[31] Kiyohara S and Kumagai Y 2025 *J. Phys. Chem. Lett.* **16** 5244

[32] Cui G, Guo Z, Ren X, Jiang Y, Jin X, Wu Y and Ju S 2025 *J. Phys. Chem. Lett.* **16** 3579

[33] Tallapally D, Wang J, Potika K and Eirinaki M 2023 *Algorithms* **16** 515

[34] Ma J, Liu B, Li K, Li C, Zhang F, Luo X and Qiao Y 2024 *Neurocomputing* **609** 128490

[35] Raissi M, Perdikaris P and Karniadakis G E 2019 *J. Comput. Phys.* **378** 686

[36] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *npj Comput. Mater.* **2** 16028

[37] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301

[38] Wu Z Q, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 *Chem. Sci.* **9** 513

[39] Hanakata P Z, Cubuk E D, Campbell D K and Park H S 2018 *Phys. Rev. Lett.* **121** 255304

[40] Zheng X L, Zheng P and Zhang R Z 2018 *Chem. Sci.* **9** 8426

[41] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 *J. Chem. Phys.* **148** 241722

[42] Zhu Z, Dong B, Guo H, Yang T and Zhang Z 2020 *Chin. Phys. B* **29** 046101

[43] Chen L H, Tran H, Batra R, Kim C and Ramprasad R 2019 *Comput. Mater. Sci.* **170** 109155

[44] Chibani S and Coudert F X 2020 *APL Mater.* **8** 080701

[45] Zhao Y, Cui Y X, Xiong Z, Jin J, Liu Z H, Dong R Z and Hu J J 2020 *ACS Omega* **5** 3596

[46] Wan Z, Wang Q D, Liu D and Liang J 2021 *Comp. Mater. Sci.* **198** 110699

[47] Liu H, Xu L, Ma Z, Li Z, Li H, Zhang Y, Zhang B and Wang L L 2023 *Mater. Today Commun.* **36** 106578

[48] Hu J, Zuo Y, Hao Y, Shu G, Wang Y, Feng M, Li X, Wang X, Sun J, Ding X, Gao Z, Zhu G and Li B 2023 *Chin. Phys. B* **32** 046301

[49] Zhang W B, Guo J, Lv X K and Zhang F C 2024 *J. Phys. Chem. Lett.* **15** 5413

[50] Liu B and Wang Y 2024 *Chin. Phys. B* **33** 084401

[51] Chen Z, Wang J, Li C, Liu B, Luo D, Min Y, Fu N and Xue Q 2024 *J. Mater. Chem. C* **12** 15444

[52] Zhang Y, Wang X, Wang Y and Zheng W 2024 *Chin. Phys. B* **33** 070702

[53] Liu H, Weng M, Huang Y, Luo Y and Zheng Y 2025 *Comput. Mater. Sci.* **255** 113853

[54] Zhang Z, Liu S, Xiong Q and Liu Y 2025 *J. Phys. Chem. Lett.* **16** 738

[55] Cao B, Dong J, Wang Z and Wang L 2025 *J. Phys. Chem. Lett.* **16** 4907

[56] Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt P S, Hinsche N F, Gjerding M N, Torelli D, Larsen P M, Riis-Jensen A C, Gath J, Jacobsen K W, Mortensen J J, Olsen T and Thygesen K S 2018 *2D Mater.* **5** 042002

[57] Gjerding M N, Taghizadeh A, Rasmussen A, Ali S, Bertoldo F, Deilmann T, Knøsgaard N R, Kruse M, Larsen A H, Manti S, Pedersen T G, Petralanda U, Skovhus T, Svendsen M K, Mortensen J J, Olsen T and Thygesen K S 2021 *2D Mater.* **8** 044002

[58] Kipf T N and Welling M 2017 arXiv e-prints arXiv:1609.02907

[59] Veličković P, Cucurull G, Casanova A, Romero A, Liò P and Bengio Y 2018 arXiv e-prints arXiv:1710.10903

[60] Chen Y T, Huang D, Zhang D X, Zeng J S, Wang N Z, Zhang H R and Yan J Y 2021 *J. Comput. Phys.* **445** 110624

[61] Rong M, Zhang D X and Wang N Z 2022 *Complex Intell. Syst.* **8** 4849

[62] Lebeda T, Aschebrock T, Sun J, Leppert L and Kümmel S 2023 *Phys. Rev. Mater.* **7** 093803

[63] Kokott S, Merz F, Yao Y, Carbogno C, Rossi M, Havu V, Rampp M, Scheffler M and Blum V 2024 *J. Chem. Phys.* **161** 024112