

Article

Symm-CGNN: Symmetry-Information-Enhanced Crystal Graph Neural Network for High-Symmetry Point Band Gap Prediction

Qihang Xu ¹ , Jian Wu ² , Xiuying Zhang ³ and Sicong Zhu ^{1,*}

¹ The State Key Laboratory for Refractories and Metallurgy, Hubei Province Key Laboratory of Systems Science in Metallurgical Process, School of Physics and Mechanics, Wuhan University of Science and Technology, Wuhan 430081, China; xqh0213@wust.edu.cn

² College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China; wujian15@nudt.edu.cn

³ School of Physics and Mechanics, Wuhan University of Technology, Wuhan 430070, China; xyzhang25@whut.edu.cn

* Correspondence: sczhu@wust.edu.cn

Abstract

Accurately characterizing the anisotropic optoelectronic properties of crystals requires determining the band gaps at specific high-symmetry points in the Brillouin zone. Relying solely on the minimum band gap is insufficient. However, although Graph Neural Networks (GNNs) offer rapid property predictions, conventional models remain trapped in a local real-space paradigm, lacking the global symmetry information necessary to differentiate these high-symmetry energy states. To address this problem, we propose the Symmetry-Information-Enhanced Crystal Graph Neural Network (Symm-CGNN). It explicitly incorporates local atomic environments with global symmetry information, including space groups, crystal systems, material density and lattice constants. The evaluation is performed on a comprehensive dataset including 3D (Materials Project) and 2D (2DMatpedia). The results demonstrate that Symm-CGNN achieves an 18% reduction in Mean Absolute Error (MAE) for high-symmetry band gap prediction compared to the baseline Crystal Graph Convolutional Neural Networks (CGCNN). This approach bridges the representational gap between local atomic coordination and macroscopic symmetry. Consequently, it provides a robust and efficient machine-learning paradigm for the high-throughput screening of materials with anisotropic optoelectronic properties.

Keywords: Symm-CGNN; high-symmetry point band gap; global symmetry information; CGCNN; anisotropic optoelectronic properties



Academic Editor: Gyaneshwar P. Srivastava

Received: 31 May 2026

Revised: 9 July 2026

Accepted: 10 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The minimum band gap is a key parameter for evaluating the optoelectronic properties of semiconductors. However, this single scalar is insufficient to fully characterize complex photoexcitation processes [1–14]. It is also necessary to investigate band-edge renormalization phenomena, which typically arise from band gap differences at various high-symmetry points in the Brillouin zone [15–18]. The optical absorption profile of a material is governed by a series of vertical interband transitions at specific k-points. Instead of being defined solely by the minimum band gap, distinct absorption peaks correspond to local energy gaps at specific high-symmetry points, such as the G, K, and M points (Figure 1a). Therefore, determining the band gaps at these high-symmetry points is necessary for accurately mapping the optical absorption spectrum.

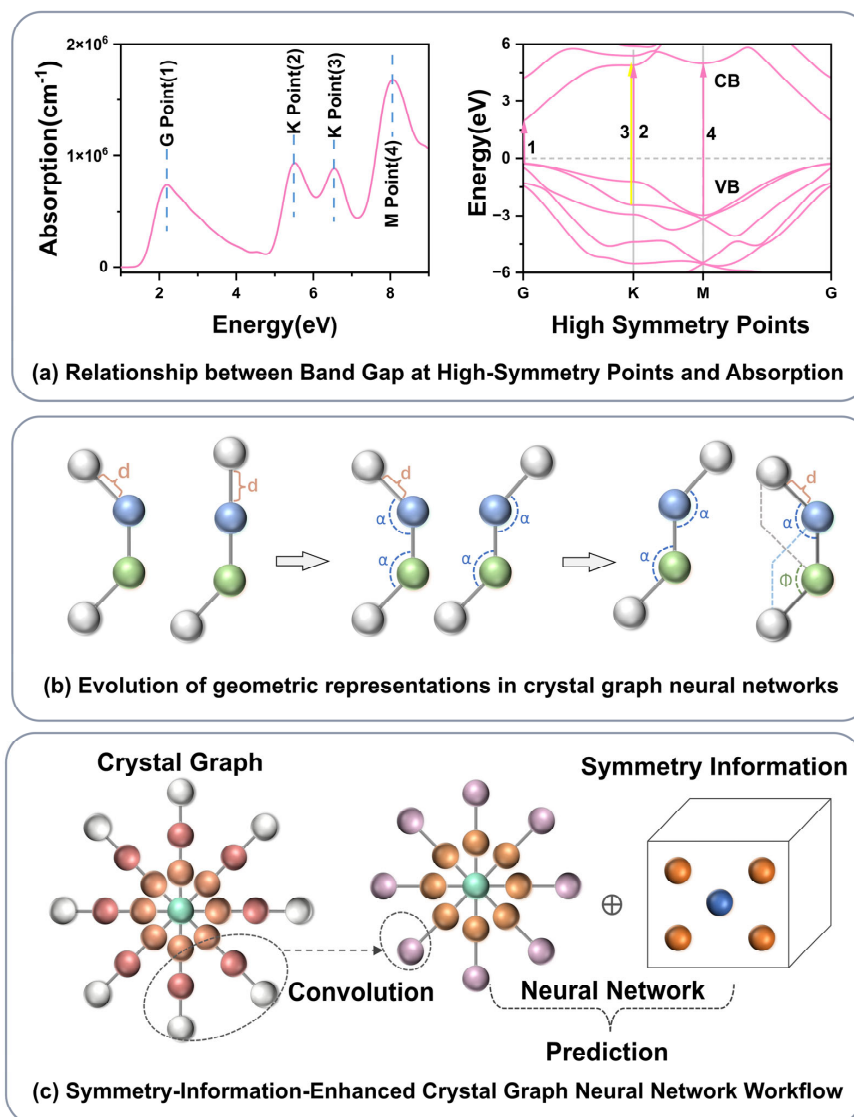


Figure 1. Physical motivation for predicting high-symmetry band gaps and the proposed deep learning architecture. (a) Schematic illustration bridging the macroscopic optical absorption spectrum (left) and the electronic band structure (right). The distinct absorption peaks directly correspond to specific vertical interband transitions (denoted as 1–4) occurring at various high-symmetry k-points (K, M, G). (b) The evolution and inherent limitations of conventional local deep learning models. (c) The workflow of the proposed Symm-CGNN.

Although standard first-principles calculations provide accurate high-symmetry band gaps [19–21], their computational cost limits their application in high-throughput screening of large-scale materials databases [22–26]. To address this limitation, deep learning techniques have emerged as efficient approaches for material property prediction. In general, such models characterize the geometric features of crystals based on local structures. Model architectures have evolved to capture increasingly complex local structures, moving from bond lengths (e.g., CGCNN, SchNet) to bond angles (e.g., MEGNet) and dihedral angles (e.g., SphereNet) (Figure 1b). Despite these advancements, they remain confined to local real-space representations. Conventional models often fail to accurately predict high-symmetry band gaps due to the absence of global symmetry information [27–33]. Consequently, these local models exhibit limitations in distinguishing electronic state differences between various high-symmetry points, which restricts their prediction accuracy [34–37].

In this work, we propose a deep learning framework, Symm-CGNN, which incorporates global symmetry information. Built upon the CGCNN architecture, this model retains local chemical environment modeling while integrating global symmetry constraints to provide a representation of electronic structures [38–40]. To integrate these global structural features, specific encoding strategies are employed [41]. Raw crystal variables are mapped into continuous vector representations to preserve essential structural relationships [42]. As shown in the workflow (Figure 1c), Symm-CGNN extracts local structural features through graph convolutions and encodes macroscopic crystal properties as symmetry information. By combining these local and global features, the model accounts for overall crystal symmetry, leading to improved prediction accuracy of band gaps at specific high-symmetry points. Notably, this framework achieves a 26% reduction in MAE for high-symmetry band gap prediction compared to the baseline CGCNN.

2. Model Architecture

Based on the CGCNN architecture, Symm-CGNN constructs graph representations directly from Crystallographic Information Files (CIFs). In this topology, atoms are represented as nodes encoded with elemental properties (Table 1), and chemical bonds within a defined cutoff radius are represented as edges encoded with interatomic distances (Figure 2a).

Table 1. Elemental properties used for atomic node feature representation.

Property	Unit	Range	Categories
Group number	-	1, 2, ..., 18	18
Period number	-	1, 2, ..., 9a	9
Electronegativity	-	0.5–4.0	10
Covalent radius	pm	25–250	10
Valence electrons	-	1, 2, ..., 12	12
First ionization energy	eV	1.3–3.3	10
Electron affinity	eV	−3–3.7	10
Block	-	s, p, d, f	4
Atomic volume	cm ³ /mol	1.5–4.3	10

To supplement the local graph, macroscopic structural properties, including crystal volume (Ω), crystal density (ρ), lattice constants ($\vec{a}$, $\vec{b}$), crystal systems, and space groups (Table 2), are extracted as global features (η). During information propagation (Figure 2b), node and edge features are processed through multiple graph convolutional layers to iteratively aggregate local structural information. A pooling operation then aggregates the updated node features into a graph-level representation, which is concatenated with the global features (η). Finally, this combined representation passes through linear layers to predict the band gap.

Table 2. The macroscopic structural properties used in Global features.

Property	Unit	Range	Categories
Lattice a	Å	2–18	11
Lattice b	Å	2–18	11
Volume	Å ³	0–5000	26
Density	g/cm ³	0–3.6	11
Lattice type	-	0, 1, ..., 6	7
Space group	-	0, 1, ..., 230	16

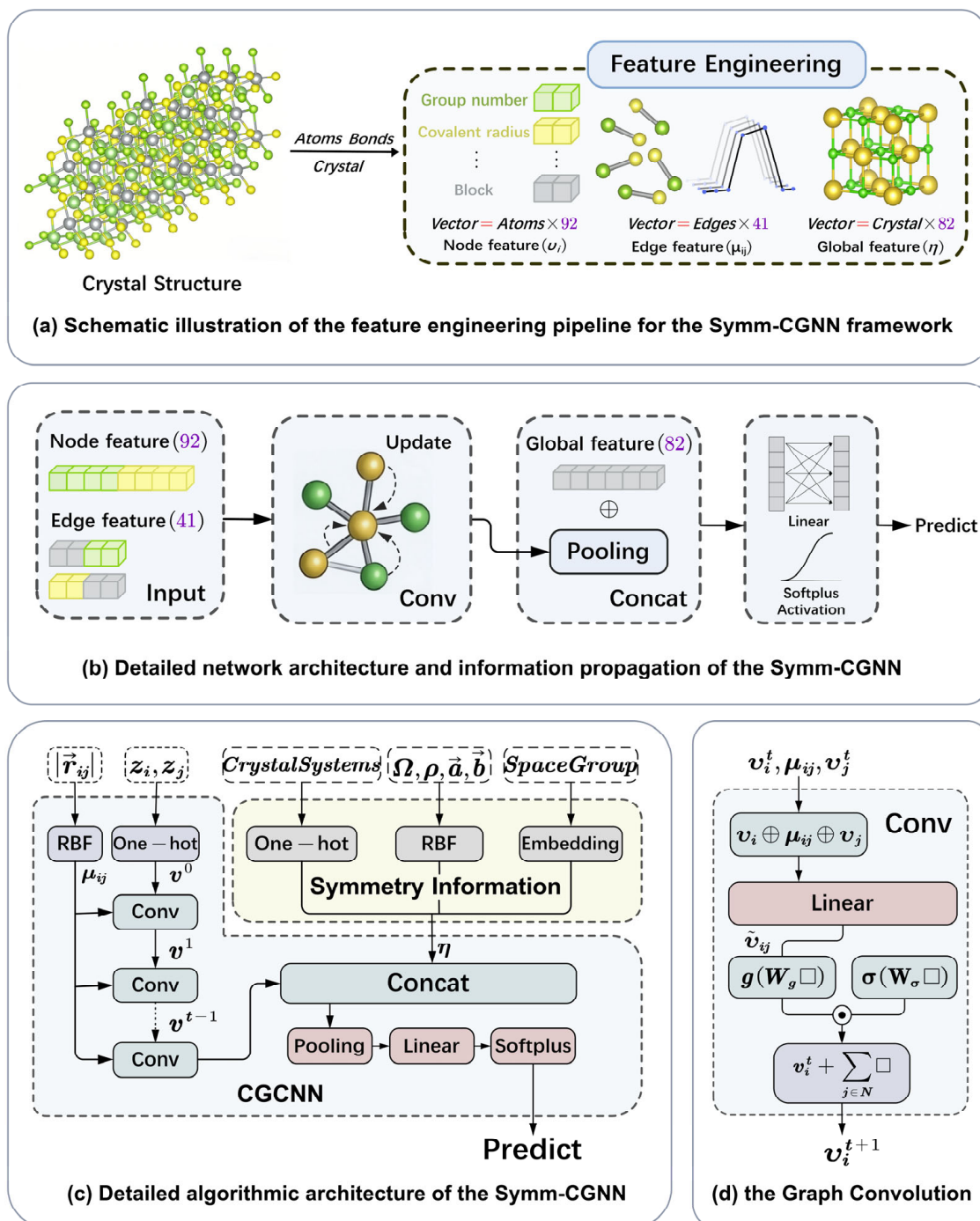


Figure 2. Comprehensive architecture and algorithmic workflow of the Symm-CGNN framework. (a) Schematic illustration of the feature engineering pipeline, mapping the raw crystal structure into explicit node (atom), edge (bond), and global (macroscopic) representations. (b) network architecture and information propagation, highlighting the extraction of local atomic environments via graph convolutions and the subsequent late-fusion of global features. (c) Detailed algorithmic schematic. Continuous spatial variables (e.g., distance, volume) and discrete crystallographic properties (e.g., space group, crystal system) are encoded via Gaussian radial basis functions (RBF) and embedding/one-hot layers to form the global context (η), which is then concatenated with the pooled local graph features. (d) Inner workings of the graph convolution block, demonstrating the local message-passing mechanism updated via gated linear operations.

The feature encoding workflow processes both discrete and continuous variables (Figure 2c). Discrete categorical features (e.g., crystal systems and space groups) are rep-

resented using one-hot encoding and learnable embedding layers. Conversely, Gaussian radial basis functions (RBFs) are used to expand continuous variables. Inspired by Phys-Net [42], RBFs are used to map both interatomic distances and continuous macroscopic features (e.g., volume and density) into high-dimensional vector representations, as expressed in Equation (1):

$$\mu_{ij} = \exp\left(-\beta_k \cdot (\exp(-|r_{ij}|) - v_k)^2\right) \quad (1)$$

where r_{ij} denotes the distance between atom i and atom j , while β_k and v_k are learnable parameters of the k -th Gaussian basis function, used to adjust the center position and distribution width of the basis functions. Through the above formulation, interatomic distances are transformed into vectors of appropriate dimensionality, serving as edge feature vectors μ_{ij} .

Subsequently, through multiple convolutional layers, each node aggregates information from itself and its neighboring nodes, thereby updating the representations of all nodes in the graph. The main formulations are given in Equations (2) and (3):

$$v_i^{t+1} = v_i^t + \sum_{j,k} \sigma\left(z_{i,j}^t W_f^t + b_f^t\right) \odot g\left(z_{i,j}^t W_f^t + b_f^t\right) \quad (2)$$

$$z_{ij}^t = v_i^t \oplus u_{ij} \oplus v_j^t \quad (3)$$

where $\odot$ denotes element-wise multiplication, W_f^t and b_f^t represent the weight matrix and bias vector of the t -th convolutional layer, respectively, σ denotes the Sigmoid function, g denotes the Softplus function, and $\oplus$ represents the concatenation of node features and edge features. Figure 2d illustrates the detailed process by which the convolutional layers aggregate feature information from each node and its neighboring nodes.

After multiple convolutional layers, a pooling layer is applied to generate the overall crystal feature vector v_c . For simplicity, normalized summation is adopted as the pooling function, as shown in Equation (4).

$$v_c = \text{SUM}(v_0^t, v_1^t, v_2^t \cdots v_N^t) \quad (4)$$

Finally, the overall feature vector v_c and the global feature vector η are combined through hidden layers to capture the complex mapping between crystal structures and their properties. An output layer is then connected to the hidden layers to predict the target properties.

3. Results

The model was trained and evaluated on a combined dataset constructed from the Materials Project (MP) [43] (comprising 3D crystal structures) and the 2DMatPedia (2DMat) [44] (comprising 2D materials) databases. Training on a unified dataset that includes both three-dimensional and two-dimensional materials not only enhances the diversity of the data distribution but also improves the model's ability to represent materials with different structural dimensionalities. The dataset spans seven major crystal systems, with the overall distribution pattern remaining consistent across both the 3D and 2D material subsets. This distribution provides diverse and representative structures in terms of both structural dimensionality and crystal symmetry, offering a wide range of training examples for the model to capture structural variations (Figure 3). The frequency distribution of the dominant space groups is detailed in Figure A2. In the selection of prediction targets, we analyzed the band structure data of all materials in the database. Based on the high-symmetry point path information contained in the data, the band gap values at each high-symmetry point were calculated, and the band gaps at the X, Y, Z and Γ points (Gap_X , Gap_Y , Gap_Z and Gap_Γ) were taken as the primary prediction targets. Prior to model training, we performed

data preprocessing to remove anomalous samples and extreme outliers. From a physical perspective, ultra-wide-bandgap materials are rare in practical optoelectronic applications and generally suffer from higher calculation uncertainty in first-principles simulations, and these extreme samples would impair the model's fitting performance for the mainstream band gap range. Accordingly, we set independent one-sided upper truncation thresholds for the four targets: 6 eV for Gap_X and Gap_Y , 7 eV for Gap_Y , 8 eV for Gap_Z . The statistical distributions of the four curated high-symmetry band gaps are visualized in Figure A3. After preprocessing, the final valid sample sizes for model training and evaluation are 10,146 (Gap_X), 10,871 (Gap_Y), 15,194 (Gap_Z), and 23,163 (Gap_T), respectively. The full workflow of data curation and model evaluation is summarized in Figure A1.

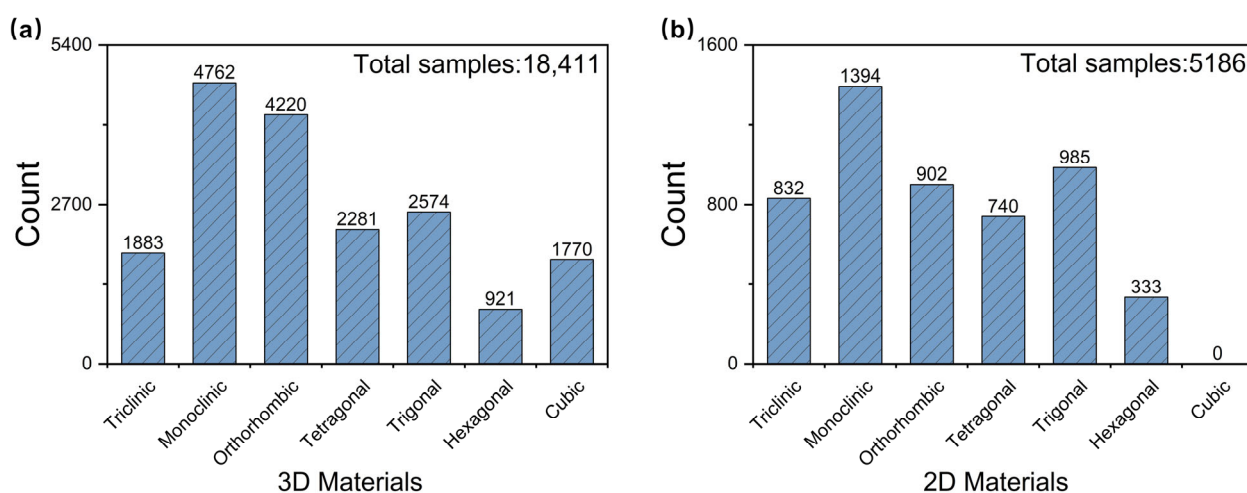


Figure 3. Crystal system distribution of materials in the dataset across different dimensionalities. (a) Sample counts of the seven major crystal systems for the 3D materials. (b) Sample counts of the crystal systems for the 2D materials.

Ablation studies are widely utilized to validate the efficacy of individual input features and quantify their respective contributions to predictive performance. To quantify the physical properties to model performance, we performed ablation studies using the Full Model (FM) as the baseline. Six ablation control groups were established by excluding individual feature categories: density (w/o Den), unit cell volume (w/o Vol), lattice constant a (w/o Lat_a), lattice constant b (w/o Lat_b), lattice type (w/o LT), and space group (w/o SG). To ensure a rigorous control of variables, all ablation configurations maintained the same model architecture, training hyperparameters, dataset splits, and random seeds as the baseline (Table 3). As shown in Table 4, the experimental results show that the Full Model (FM) achieves the lowest MAE across all targets, indicating that all selected features contribute to the model's performance. Specifically, removing the space group (w/o SG) has the most significant impact on Gap_X , increasing the MAE from 0.35 to 0.48. For Gap_Z , omitting the unit cell volume (w/o Vol) leads to the highest error (MAE of 0.58). Other features, such as density and lattice constants, also exhibit varying degrees of importance depending on the specific band gap being predicted. This indicates that integrating these diverse physical and structural properties is beneficial for improving overall prediction accuracy. Physically, the observed trends are consistent with the symmetry-driven origin of electronic energy bands, where symmetry features like space groups and lattice types govern band degeneracy and dispersion relations at high-symmetry points. As these features act as primary structural determinants of the electronic bandgap, their omission inevitably leads to a visible decline in predictive accuracy.

Table 3. Hyperparameter configurations for the neural network model.

Category	Hyperparameter	Value
Architecture	Number of convolutional layers	3
	Number of hidden layers	3
Training	Optimizer	Adam
	Learning rate	0.01
	Batch size	64
	Epochs	500
Regularization	Dropout fraction	0.1
	Weight decay	1×10^{-4}

Table 4. The MAE comparison of different ablation groups on bandgap predictions.

Group	Ablation Study MAE Comparison			
	Gap_X	Gap_Y	Gap_Z	Gap_Γ
FM	0.35	0.48	0.52	0.48
w/o Den	0.44	0.51	0.55	0.53
w/o Lat_a	0.36	0.53	0.56	0.51
w/o Lat_b	0.37	0.51	0.56	0.52
w/o LT	0.36	0.50	0.57	0.50
w/o Vol	0.44	0.52	0.58	0.52
w/o SG	0.48	0.51	0.57	0.52

A 10-fold cross-validation was conducted to evaluate the model's generalization capability. The full validation MAE convergence curves for all four high-symmetry band gaps during cross-validation are provided in Figures A4–A7. Across the 10 independent folds, the Symm-CGNN framework yields lower Mean Absolute Errors (MAEs) for all high-symmetry points compared to the baseline CGCNN model (Figure 4). Specifically, the 10-fold average MAEs achieved by the Symm-CGNN are 0.3491 (Gap_X), 0.4827 (Gap_Y), 0.5227 (Gap_Z), and 0.4767 (Gap_Γ), which are lower than the corresponding CGCNN averages of 0.4276, 0.5671, 0.6179, and 0.5291. Furthermore, Symm-CGNN exhibits narrower error variances across the 10 folds, indicating improved prediction stability across different data splits. These results validate our core hypothesis: while conventional local graph convolutions struggle to differentiate complex electronic states, the injection of macroscopic crystal symmetry features (e.g., space groups and lattice parameters) successfully forcibly guides the network. This global-local fusion enables the Symm-CGNN to capture the underlying physical constraints of the Brillouin zone.

The predictive performance of Symm-CGNN was evaluated on independent test sets and compared against three baseline models: CGCNN, MEGNet, and ALIGNN. As shown in Table 5, where the reduction percentages represent the relative decrease in MAE of Symm-CGNN compared to each baseline model, Symm-CGNN consistently achieves lower MAE values across all four target properties. These relative error reductions range from 2% to 31%. The improvement is particularly notable for Gap_X , where Symm-CGNN reduces the prediction error by more than 25% against MEGNet and ALIGNN. For the other three targets (Gap_Y , Gap_Z , and Gap_Γ), the model maintains a steady advantage, with error reductions typically ranging between 7% and 16%. These results indicate that incorporating structural symmetry features improves band gap prediction accuracy across various high-symmetry directions. Specifically, Symm-CGNN yields a more substantial performance gain over CGCNN, which encodes no explicit crystallographic symmetry information. In comparison, the improvement is relatively moderate against MEGNet and ALIGNN, likely because these two models already integrate local geometric descriptors

(such as bond angles) that implicitly capture certain symmetry constraints. This trend further underscores the critical role of symmetry information in predicting band gaps at high-symmetry points.

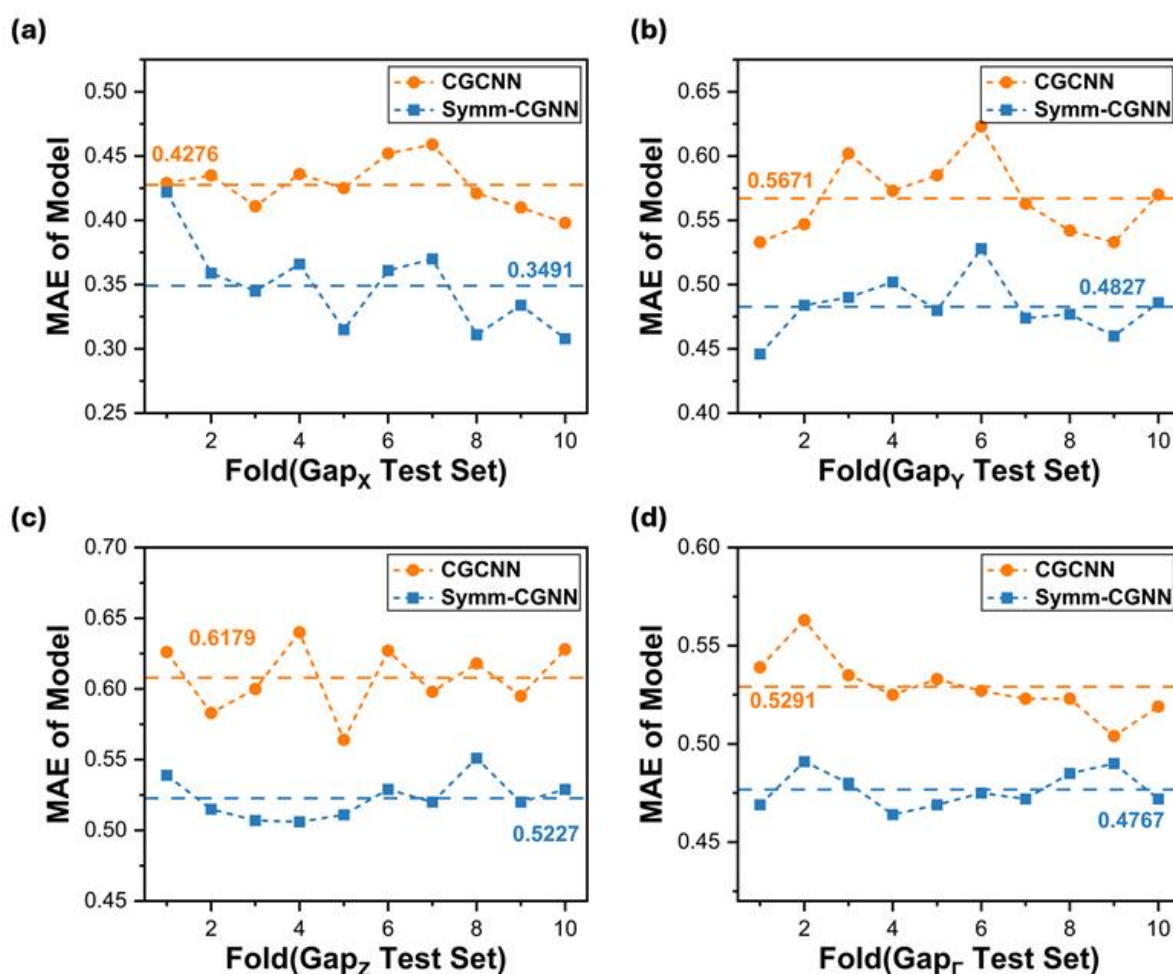


Figure 4. Predictive performance comparison under 10-fold cross-validation. (a) MAE for the Gap_X prediction across 10 independent data folds. (b) MAE for Gap_Y the prediction across 10 independent data folds. (c) MAE for the Gap_Z prediction across 10 independent data folds. (d) MAE for the Gap_Γ prediction across 10 independent data folds.

Table 5. The table summarizes the MAE of Symm-CGNN and three baseline models. The reduction percentage represents the relative MAE decrease of Symm-CGNN with respect to each corresponding baseline model.

Target	Symm-CGNN	CGCNN		MEGNet		ALIGNN	
		MAE	Reduction	MAE	Reduction	MAE	Reduction
Gap_X (eV)	0.35 ± 0.03	0.43	18%	0.51	31%	0.48	27%
Gap_Y (eV)	0.48 ± 0.02	0.57	16%	0.55	13%	0.52	8%
Gap_Z (eV)	0.52 ± 0.01	0.62	16%	0.59	12%	0.56	7%
Gap_Γ (eV)	0.48 ± 0.01	0.53	9%	0.56	14%	0.49	2%

To evaluate the model's predictive performance across all target high-symmetry band gaps, we first analyze Gap_X as a representative case. The baseline CGCNN yields an R^2 of 0.70 with a prediction error envelope of $\Delta = 0.70$ eV (Figure 5a). In comparison, Symm-CGNN achieves an R^2 of 0.83 and a narrower envelope of $\Delta = 0.60$ eV, showing tighter clustering of the predicted values along the diagonal (Figure 5c). This improvement

is also reflected in the residual error distributions. The baseline CGCNN has an error distribution with a mean (μ) of -0.01483 and a standard deviation (σ) of 0.662 (Figure 5b). For Symm-CGNN, the σ decreases to 0.493 (Figure 5d), indicating a narrower error spread and reduced prediction variance.

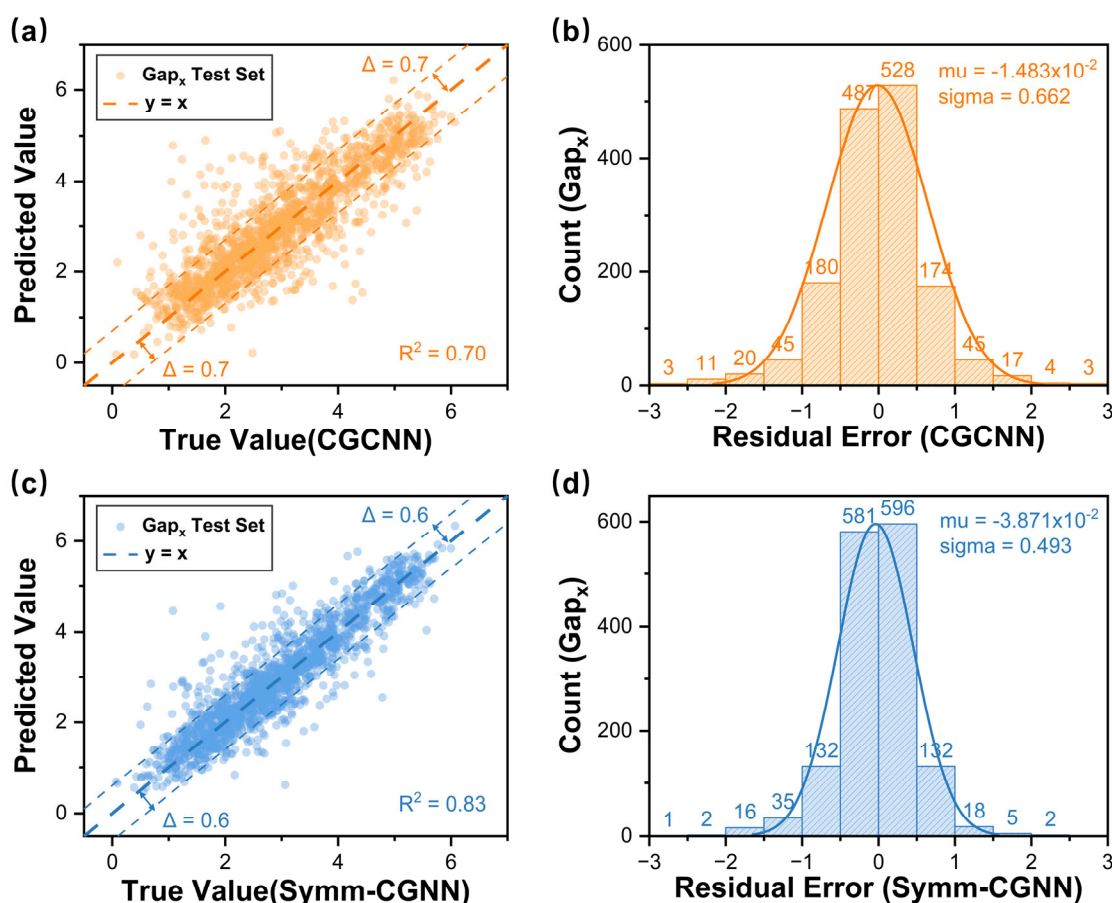


Figure 5. Comparison of prediction accuracy and residual error distribution for Gap_X between CGCNN and Symm-CGNN. (a,b) predicted versus true values and the corresponding residual error distribution for the CGCNN. (c,d) predicted versus true values and the corresponding residual error distribution for the Symm-CGNN.

For Gap_Y , Symm-CGNN demonstrates a consistent improvement. The CGCNN baseline obtains an R^2 of 0.66 with an error envelope of $\Delta = 0.9$ eV (Figure 6a). In contrast, Symm-CGNN increases the R^2 to 0.81 and narrows the error envelope to $\Delta = 0.7$ eV, demonstrating a tighter concentration of predicted values along the diagonal (Figure 6c). Regarding residual distributions, the standard deviation of the prediction errors drops from 0.866 for CGCNN to 0.655 for Symm-CGNN, while the mean residual remains close to zero for both models (Figure 6b,d).

In terms of Gap_Z , CGCNN yields an R^2 of 0.77 and an error envelope of $\Delta = 0.95$ eV (Figure 7a), whereas Symm-CGNN achieves an R^2 of 0.83 with a reduced error envelope of $\Delta = 0.75$ eV (Figure 7c). The residual standard deviation decreases from 0.851 to 0.716 upon incorporating symmetry information, indicating the model's enhanced prediction consistency (Figure 7b,d).

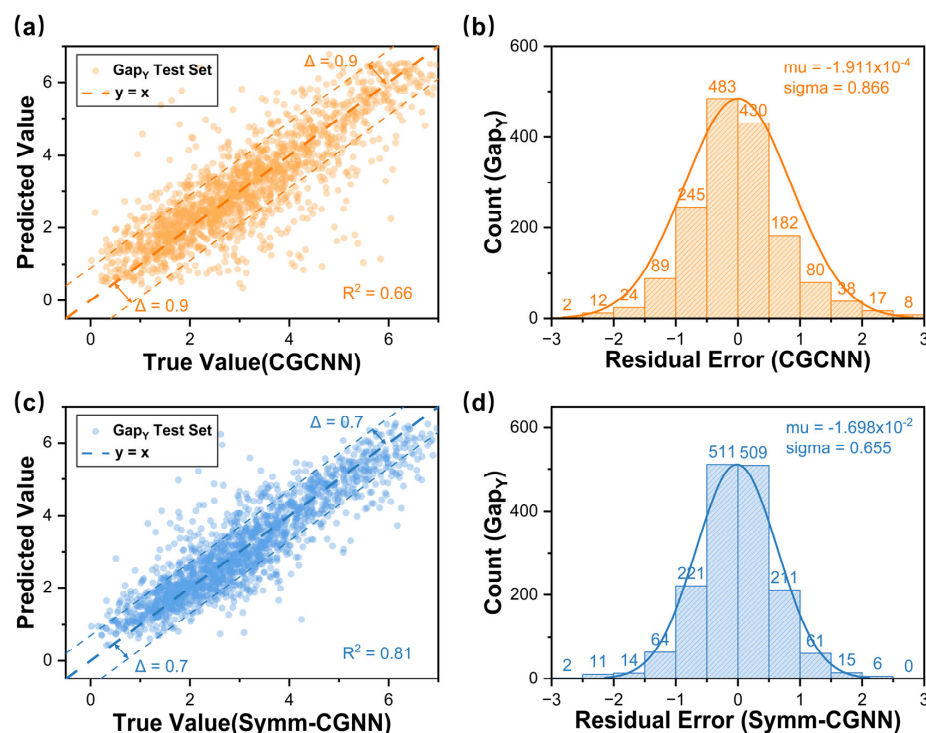


Figure 6. Comparison of prediction accuracy and residual error distribution for Gap_Y between CGCNN and Symm-CGNN. (a,b) predicted versus true values and the corresponding residual error distribution for the CGCNN. (c,d) predicted versus true values and the corresponding residual error distribution for the Symm-CGNN.

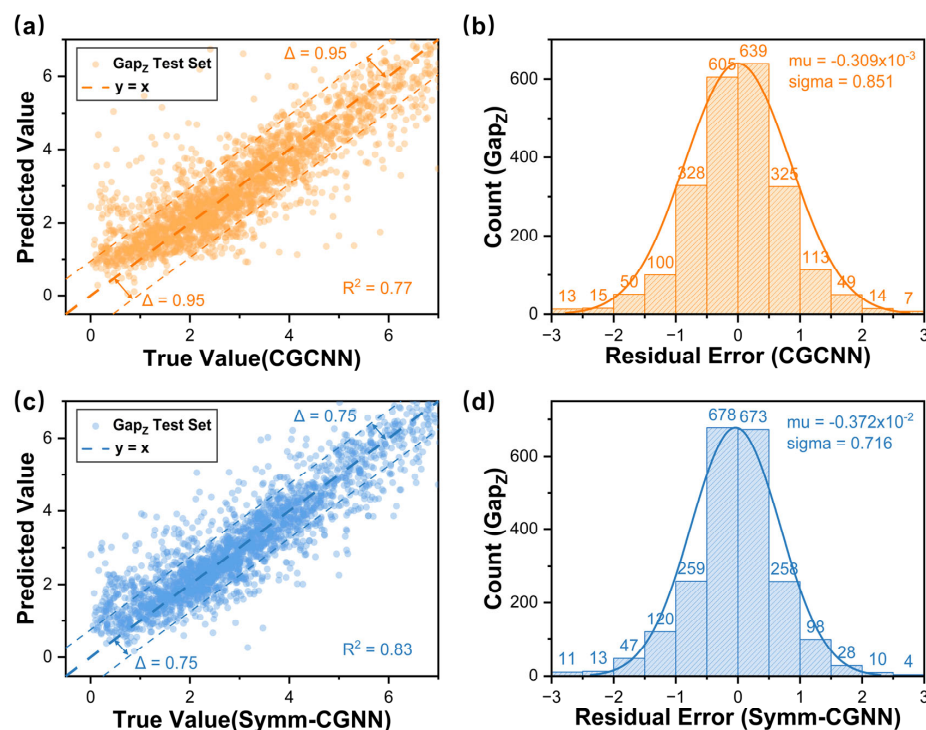


Figure 7. Comparison of prediction accuracy and residual error distribution for Gap_Z between CGCNN and Symm-CGNN. (a,b) predicted versus true values and the corresponding residual error distribution for the CGCNN. (c,d) predicted versus true values and the corresponding residual error distribution for the Symm-CGNN.

For Gap_T , which possesses the largest sample size, CGCNN yields and R^2 of 0.64 with an error envelope of $\Delta = 0.7$ eV (Figure 8a). Symm-CGNN improves the R^2 to 0.78 and narrows the error envelope (Figure 8c). The standard deviation of the residual distribution falls from 0.867 to 0.674, indicating that incorporating global symmetry features consistently reduces prediction variance across the evaluated high-symmetry points (Figure 8b,d).

From the MAE results, it can be observed that the Symm-CGNN consistently outperforms the CGCNN across different prediction targets. This indicates a better fit to the data and lower prediction errors, demonstrating a significant overall improvement in predictive capability.

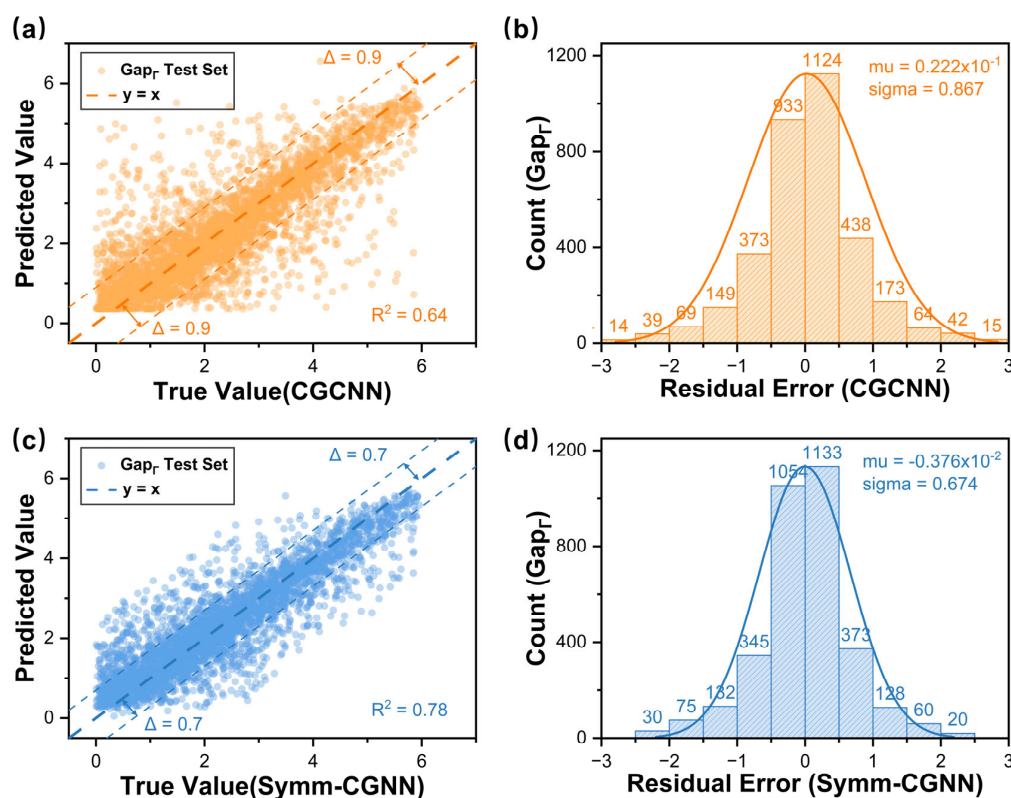


Figure 8. Comparison of prediction accuracy and residual error distribution for Gap_T between CGCNN and Symm-CGNN: (a,b) predicted versus true values and the corresponding residual error distribution for the CGCNN; (c,d) predicted versus true values and the corresponding residual error distribution for the Symm-CGNN.

4. Discussion

Overall, our Symm-CGNN model delivers remarkably favorable prediction performance for band gaps at high-symmetry points, outperforming conventional local graph neural network baselines across all evaluation metrics. Unlike conventional local-graph models constrained by short-range atomic environments, Symm-CGNN incorporates macroscopic parameters—such as volume, density, lattice constants, and space groups—to enhance prediction accuracy for symmetry-dependent band gaps at high-symmetry points. We attribute this performance improvement to the physical nature of reciprocal space, where electronic energy states at specific k -points are governed by the crystal's long-range periodicity and global symmetry operations (such as space groups) that cannot be fully characterized by local chemical environments alone. Crystals with similar local atomic arrangements can form distinct macroscopic lattices; consequently, without global symmetry information, purely local models may fail to distinguish these structural differences, leading to deviations in reciprocal-space property predictions. The introduced space groups and lattice constants

provide Symm-CGNN with essential geometric boundary conditions and physical constraints, supporting more reliable predictions of band gaps at high-symmetry points.

Building on the effectiveness of this symmetry-integration strategy, the model's performance could be further enhanced in future work. For instance, applying this framework to models that already incorporate local geometric descriptors (such as bond and dihedral angles in ALIGNN), or adopting the design principles of equivariant neural networks, represents a promising path toward higher accuracy [45,46]. Additionally, expanding the training dataset scale in the context of high-throughput materials research is expected to further improve the model's prediction accuracy and generalization capability.

5. Conclusions

This study addresses the challenge of predicting band gaps at high-symmetry points in crystalline materials. To overcome this issue, we built the Symm-CGNN model based on the CGCNN framework, which incorporates the global geometric and symmetry features of crystals and encodes them into high-dimensional representations. The architecture combines local structural awareness with global geometric constraints. Evaluated on a dataset combining the MP and 2DMat databases, Symm-CGNN achieves a 26% reduction in prediction MAE relative to the CGCNN baseline. In summary, Symm-CGNN improves the predictive accuracy of high-symmetry band gaps, addressing the limitations of conventional models in distinguishing band gaps across the Brillouin zone. By reducing prediction errors in both 3D and 2D material tasks, this model provides a tool for predicting electronic structure properties and offers a methodological framework for subsequent studies on band-edge renormalization, photoexcited transitions, and optoelectronic properties.

Author Contributions: Conceptualization, S.Z.; methodology, X.Z.; validation, Q.X.; formal analysis, Q.X.; investigation, Q.X.; resources, S.Z.; data Q.X.; writing—original draft preparation, Q.X.; writing—review and editing, J.W.; visualization, X.Z.; supervision, S.Z.; project administration, X.Z.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China under Grants (No. 11704291, No.12174296 and No.52475210), and Wuhan University of Science and Technology Student Innovation and Entrepreneurship Project (No.: Z25070).

Data Availability Statement: The datasets used in this study were obtained from the Materials Project and 2DMatpedia.

Acknowledgments: The authors would like to thank the supported by High-Performance Computing Center of Wuhan University of Science and Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Symm-CGNN	Symmetry-Information-Enhanced Crystal Graph Neural Network
CGCNN	Crystal Graph Convolutional Neural Network
MAE	Mean Absolute Error
Gap _X	The electronic band gap at the X high-symmetry point
Gap _Y	The electronic band gap at the Y high-symmetry point
Gap _Z	The electronic band gap at the Z high-symmetry point
Gap _Γ	The electronic band gap at the Γ high-symmetry point

Appendix A

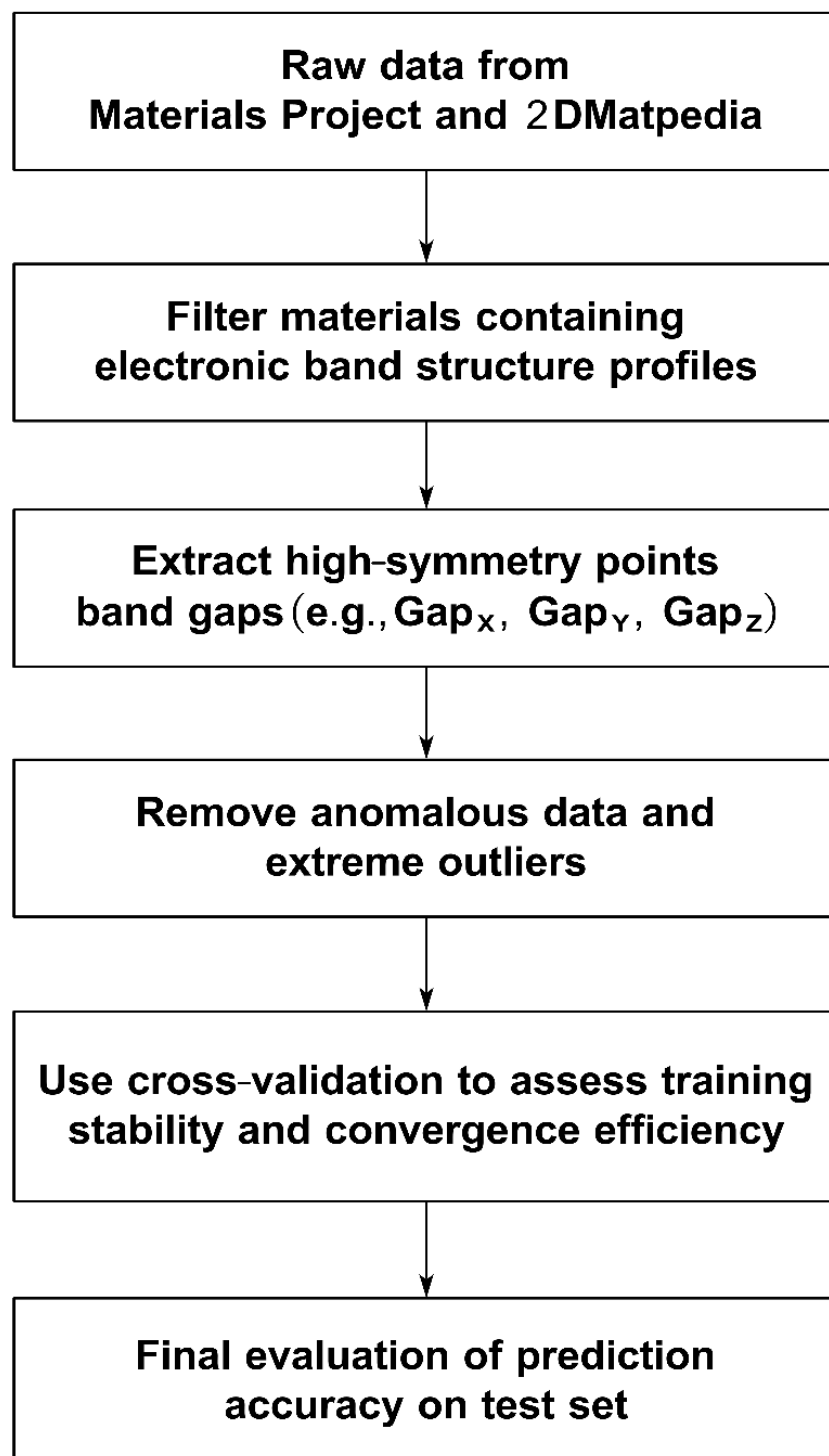


Figure A1. Comprehensive workflow of the data curation and model evaluation. The process begins with raw data from the Materials Project and 2DMatpedia databases. Subsequent data screening involves filtering materials with complete electronic band structures, extracting specific high-symmetry band gaps, and rigorously removing extreme outliers. The curated dataset is then subjected to a dual-phase assessment: cross-validation to verify training stability and convergence efficiency, followed by a final evaluation on an independent test set to ascertain the ultimate prediction accuracy.

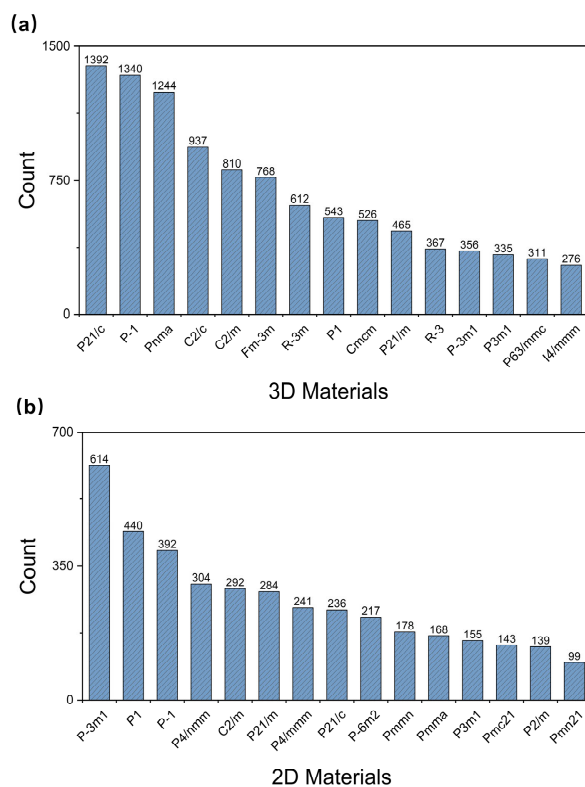


Figure A2. Distribution of the top 15 most frequent space groups in the dataset. (a) Space group frequency distribution for 3D materials; (b) Space group frequency distribution for 2D materials.

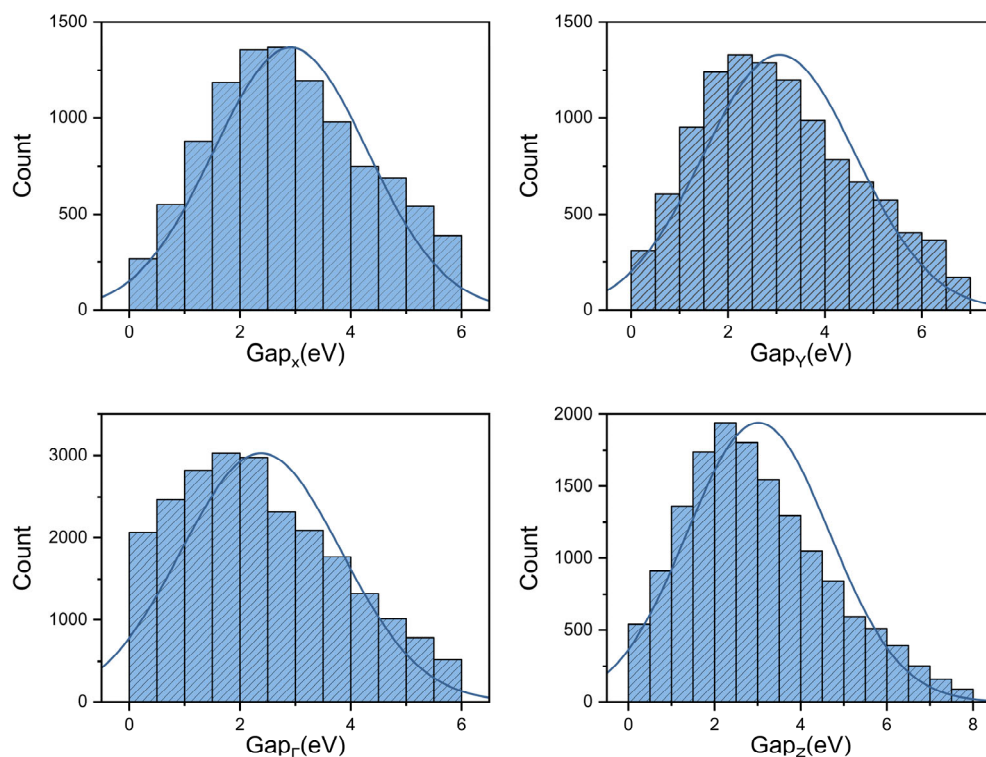


Figure A3. Statistical distribution of target high-symmetry band gaps in the curated dataset. The histograms display the frequency distributions of local energy gaps extracted at various critical high-symmetry points in the Brillouin zone. The solid blue curves indicate the overlaid continuous probability density estimations.

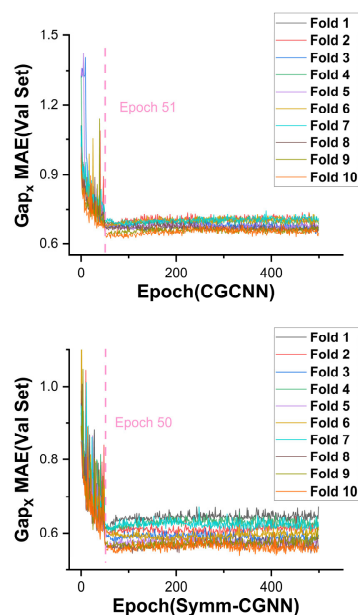


Figure A4. Validation MAE curves over 500 epochs for Gap_X predictions. The upper and lower plots display the convergence behaviors of the baseline CGCNN and the proposed Symm-CGNN frameworks during the 10-fold cross-validation, respectively. The distinct colored lines represent the 10 independent folds, while the vertical dashed lines indicate the approximate epochs where the models achieve stable convergence.

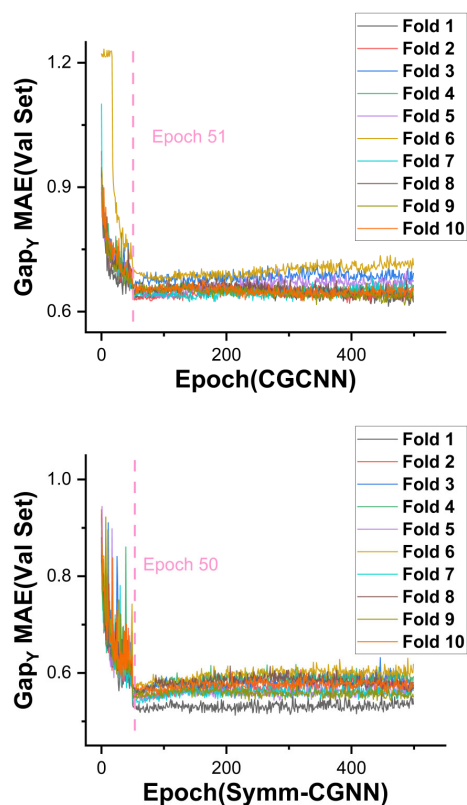


Figure A5. Validation MAE curves over 500 epochs for Gap_Y predictions. The upper and lower plots display the convergence behaviors of the baseline CGCNN and the proposed Symm-CGNN frameworks during the 10-fold cross-validation, respectively. The distinct colored lines represent the 10 independent folds, while the vertical dashed lines indicate the approximate epochs where the models achieve stable convergence.

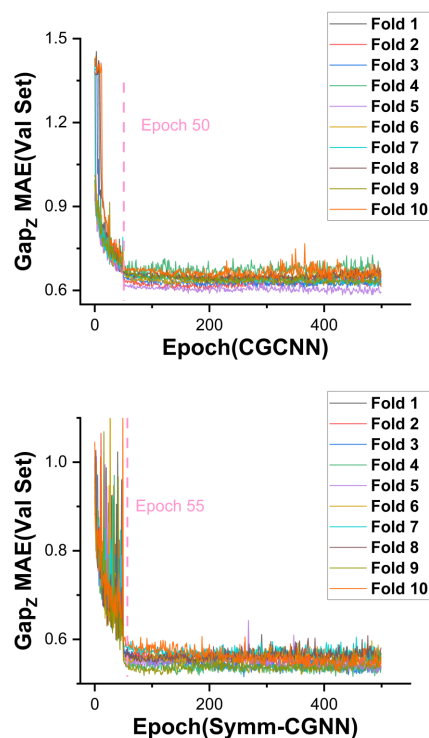


Figure A6. Validation MAE curves over 500 epochs for Gap_Z predictions. The upper and lower plots display the convergence behaviors of the baseline CGCNN and the proposed Symm-CGNN frameworks during the 10-fold cross-validation, respectively. The distinct colored lines represent the 10 independent folds, while the vertical dashed lines indicate the approximate epochs where the models achieve stable convergence.

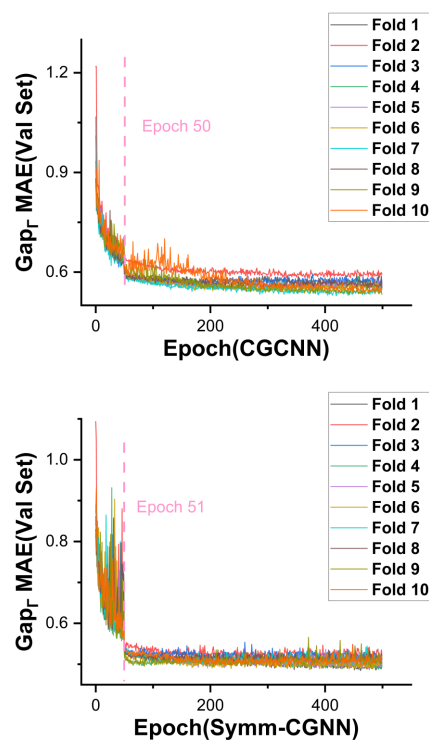


Figure A7. Validation MAE curves over 500 epochs for Gap_r predictions. The upper and lower plots display the convergence behaviors of the baseline CGCNN and the proposed Symm-CGNN frameworks during the 10-fold cross-validation, respectively. The distinct colored lines represent the 10 independent folds, while the vertical dashed lines indicate the approximate epochs where the models achieve stable convergence.

References

1. Szymanski, N.J.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.; Cubuk, E.D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **2023**, *624*, 86–91. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z.; Shysheya, A.; Crabbé, J.; Ueda, S.; et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624–632. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Liang, Z.; Zhang, Y.; Xu, H.; Chen, W.; Liu, B.; Zhang, J.; Zhang, H.; Wang, Z.; Kang, D.-H.; Zeng, J.; et al. Homogenizing out-of-plane cation composition in perovskite solar cells. *Nature* **2023**, *624*, 557–563. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Liu, J.; He, Y.; Ding, L.; Zhang, H.; Li, Q.; Jia, L.; Yu, J.; Lau, T.W.; Li, M.; Qin, Y.; et al. Perovskite/silicon tandem solar cells with bilayer interface passivation. *Nature* **2024**, *635*, 596–603. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Zhou, P.; Navid, I.A.; Ma, Y.; Xiao, Y.; Wang, P.; Ye, Z.; Zhou, B.; Sun, K.; Mi, Z. Solar-to-hydrogen efficiency of more than 9% in photocatalytic water splitting. *Nature* **2023**, *613*, 66–70. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Si, Z.; Duan, Y.; Lv, S.; Liu, S.; Yu, K.; Gan, L.; Ran, R.; Wu, X.; Lv, W.; Zhou, D.; et al. Monolithic photocatalyst for overall water splitting based on gaseous water harvest. *J. Colloid Interface Sci.* **2025**, *700*, 138579. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Hung, N.T.; Okabe, R.; Chottrattanapituk, A.; Li, M. Universal Ensemble-Embedding Graph Neural Network for Direct Prediction of Optical Spectra from Crystal Structures. *Adv. Mater.* **2024**, *36*, 2409175. [\[CrossRef\]](#)
9. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2018**, *2*, 0121. [\[CrossRef\]](#)
10. Polman, A.; Knight, M.; Garnett, E.C.; Ehrler, B.; Sinke, W.C. Photovoltaic materials: Present efficiencies and future challenges. *Science* **2016**, *352*, aad4424. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Chen, X.; Shen, S.; Guo, L.; Mao, S.S. Semiconductor-based photocatalytic hydrogen generation. *Chem. Rev.* **2010**, *110*, 6503–6570. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Takata, T.; Jiang, J.; Sakata, Y.; Nakabayashi, M.; Shibata, N.; Nandal, V.; Seki, K.; Hisatomi, T.; Domen, K. Photocatalytic water splitting with a quantum efficiency of almost unity. *Nature* **2020**, *581*, 411–414. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Stranks, S.D.; Hoyer, R.L.Z.; Di, D.; Friend, R.H.; Deschler, F. The physics of light emission in halide perovskite devices. *Adv. Mater.* **2018**, *31*, 1803336. [\[CrossRef\]](#)
14. Karlsson, M.; Yi, Z.; Reichert, S.; Luo, X.; Lin, W.; Zhang, Z.; Bao, C.; Zhang, R.; Bai, S.; Zheng, G.; et al. Mixed halide perovskites for spectrally stable and high-efficiency blue light-emitting diodes. *Nat. Commun.* **2021**, *12*, 361. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Zacharias, M.; Volonakis, G.; Giustino, F.; Even, J. Anharmonic electron-phonon coupling in ultrasoft and locally disordered perovskites. *npj Comput. Mater.* **2023**, *9*, 153. [\[CrossRef\]](#)
16. Wang, L.; Nughays, R.; Yin, J.; Shih, C.-H.; Guo, T.-F.; Mohammed, O.F.; Chergui, M. Band gap renormalization at different symmetry points in perovskites. *ACS Photonics* **2024**, *11*, 2273–2281. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Marini, A.; Poncé, S.; Gonze, X. Many-body perturbation theory approach to the electron-phonon interaction with density-functional theory as a starting point. *Phys. Rev. B* **2015**, *91*, 224310. [\[CrossRef\]](#)
18. Giustino, F. Electron-phonon interactions from first principles. *Rev. Mod. Phys.* **2017**, *89*, 015003. [\[CrossRef\]](#)
19. Borlido, P.; Aull, T.; Huran, A.W.; Tran, F.; Marques, M.A.L.; Botti, S. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **2019**, *15*, 5069–5079. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Riebesell, J.; Goodall, R.E.A.; Benner, P.; Chiang, Y.; Deng, B.; Ceder, G.; Astra, M.; Lee, A.A.; Jain, A.; Persson, K.A. Matbench Discovery—A framework to evaluate machine learning crystal stability predictions. *arXiv* **2023**, arXiv:2308.14920.
21. Unke, O.T.; Chmiela, S.; Sauceda, H.E.; Gastegger, M.; Poltavsky, I.; Schütt, K.T.; Tkatchenko, A.; Müller, K.-R. Machine learning force fields. *Chem. Rev.* **2021**, *121*, 10142–10186. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Chen, C.; Ong, S.P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2022**, *2*, 718–728. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Park, Y.J.; Yoon, S.; Jerng, S.E. Machine learning of metal-organic framework design for carbon dioxide capture and utilization. *J. CO₂ Util.* **2024**, *89*, 102941. [\[CrossRef\]](#)
24. Merchant, A.; Batzner, S.; Schoenholz, S.S.; Aykol, M.; Cheon, G.; Cubuk, E.D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Tan, H.; Teng, Y.; Shan, G. High throughput computational screening and interpretable machine learning for iodine capture of metal-organic frameworks. *npj Comput. Mater.* **2025**, *11*, 115. [\[CrossRef\]](#)
26. Boyce, B.; Dingreville, R.; Desai, S.; Walker, E.; Shilt, T.; Bassett, K.L.; Wixom, R.R.; Stebner, A.P.; Arroyave, R.; Hattrick-Simpers, J.; et al. Machine learning for materials science: Barriers to broader adoption. *Matter* **2023**, *6*, 1320–1323. [\[CrossRef\]](#)
27. Xie, T.; Grossman, J.C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Schütt, K.; Kindermans, P.J.; Sauceda Felix, H.E.; Chmiela, S.; Tkatchenko, A.; Schnet, K.-R.M. A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

29. Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564–3572. [\[CrossRef\]](#)
30. Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. *arXiv* **2020**, arXiv:2003.03123.
31. Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185. [\[CrossRef\]](#)
32. Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 6790–6802.
33. Liu, Y.; Wang, L.; Liu, M.; Lin, Y.; Zhang, X.; Oztekin, B.; Ji, S. Spherical message passing for 3d molecular graphs. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Kigali, Rwanda, 25–29 April 2022.
34. Petković, M.; Luna, J.M.V.; Dinne, E.B.; Menkovski, V.; Calero, S. Symmetry-informed graph neural networks for carbon dioxide isotherm and adsorption prediction in aluminum-substituted zeolites. *J. Mater. Chem. A* **2025**, *13*, 37018–37030. [\[CrossRef\]](#)
35. Wang, Z.; Wang, C.; Zhao, S.; Du, S.; Xu, Y.; Gu, B.-L.; Duan, W. Symmetry-adapted graph neural networks for constructing molecular dynamics force fields. *Sci. China Phys. Mech. Astron.* **2021**, *64*, 117211. [\[CrossRef\]](#)
36. Ye, D. Solving the Schrödinger Equation in Higher Dimensions with Physics-Informed Neural Networks. Master's Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2024.
37. Lee, W.S.; Kwok, I.H.; Leong, K.I.; Chau, C.K.A.; Sio, K.C. Symmetry-Constrained Multi-Scale Physics-Informed Neural Networks for Graphene Electronic Band Structure Prediction. *arXiv* **2025**, arXiv:2508.10718.
38. Louis, S.Y.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; Hu, J. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18141–18148. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Jin, L.; Du, Z.; Shu, L.; Cen, Y.; Xu, Y.; Mei, Y.; Zhang, H. Transformer-generated atomic embeddings to enhance prediction accuracy of crystal properties with machine learning. *Nat. Commun.* **2025**, *16*, 1210. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Liao, Y.L.; Wood, B.; Das, A.; Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024; pp. 39282–39309.
41. Cheng, J.; Zhang, C.; Dong, L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater.* **2021**, *2*, 92. [\[CrossRef\]](#)
42. Unke, O.T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002. [\[CrossRef\]](#)
44. Zhou, J.; Shen, L.; Costa, M.D.; Persson, K.A.; Ong, S.P.; Huck, P.; Lu, Y.; Ma, X.; Chen, Y.; Tang, H.; et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **2019**, *6*, 86. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Yang, Z.; Liu, X.; Zhang, X.; Huang, P.; Novoselov, K.S.; Shen, L. Modeling crystal defects using defect informed neural networks. *npj Comput. Mater.* **2025**, *11*, 229. [\[CrossRef\]](#)
46. Li, Y.; Wu, Y.; Han, Y.; Lyu, Q.; Wu, H.; Zhang, X.; Shen, L. Local environment interaction-based machine learning framework for predicting molecular adsorption energy. *J. Mater. Inform.* **2024**, *4*, 4. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.