ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Graph Comparison of Molecular Crystals in Band Gap Prediction Using Neural Networks

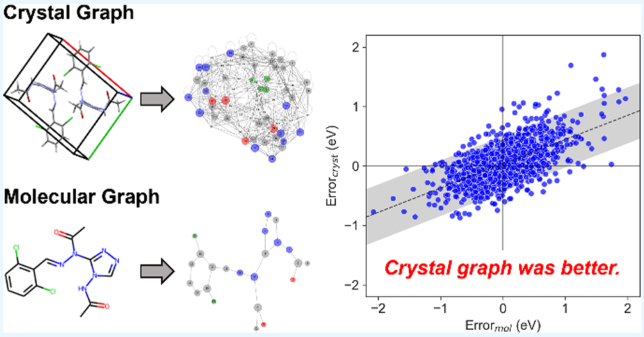
Takuya Taniguchi,* Mayuko Hosokawa, and Toru Asahi

Cite This: *ACS Omega* 2023, 8, 39481−39489

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In material informatics, the representation of the material structure is fundamentally essential to obtaining better prediction results, and graph representation has attracted much attention in recent years. Molecular crystals can be graphically represented in molecular and crystal representations, but a comparison of which representation is more effective has not been examined. In this study, we compared the prediction accuracy between molecular and crystal graphs for band gap prediction. The results showed that the prediction accuracies using crystal graphs were better than those obtained using molecular graphs. While this result is not surprising, error analysis quantitatively evaluated that the error of the crystal graph was 0.4 times that of the molecular graph with moderate correlation. The novelty of this study lies in the comparison of molecular crystal representations and in the quantitative evaluation of the contribution of crystal structures to the band gap.



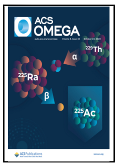
■ INTRODUCTION

Material informatics (MI) constitutes a compelling research subject, attracting significant attention across academic and industrial landscapes.[1−3] The application of MI is becoming widespread in various fields, typically in polymers and inorganics. In most cases of MI, a set of structural information and target properties is required as prior knowledge for model supervision. While a target variable can be defined relatively straightforwardly depending on the purposes, structural information can be represented in various ways. For example, conventional representations of molecules are fingerprint vectors, such as extended connectivity fingerprints.[4] Many types of descriptors have been developed,[5] and the choice of the descriptor influences the predictive performance of a task to be solved.

In recent years, graph representation has received a great deal of attention. Molecular structures can be represented as graph data consisting of nodes and edges, and graph neural networks (GNNs) can handle regression and classification tasks.[6−9] Recent studies on inorganic crystals also have successfully used graphs for inorganic crystal structures.[10−14] Graph-based approaches have been applied not only to molecules and bulk inorganic materials but also to metal−organic frameworks, two-dimensional materials, and even molecular dynamics simulations.[15−17] Furthermore, model architectures such as ALIGNN and M3GNet, which incorporate angle information, have been developed to improve prediction accuracy.[13,18]

In contrast to inorganic materials and polymers, MI research on molecular crystals has made little progress.[19−21] This may be due to the lack of databases linking the structure and properties of the molecular crystals. For example, the Cambridge Structural Database (CSD) and Crystallography Open Database (COD) are useful platforms of molecular crystal structures,[22,23] but they do not include information on physicochemical properties. Therefore, constructing structure−property data sets requires human labor to collect physical properties from publications, experiments, and quantum chemical calculations. Olsthoorn et al. constructed a large database of electronic band gaps (N = 12,500) that relates molecular crystal structures to band gaps obtained from quantum chemical calculations.[24] The band gap is the energy difference between the conduction and valence bands and is related to the optoelectronic properties. They also performed a regression analysis using crystal graph and achieved a mean absolute error (MAE) of 0.388 eV by the ensemble model of the smooth overlap of atomic positions (SOAP)[25] and a kind of GNN, SchNet.[6] It is expected that such database construction and MI-aided material screening will be helpful for the development of novel molecular crystals, as evidenced by some recent works.[26−28]

Received: July 20, 2023
Accepted: October 3, 2023
Published: October 16, 2023



https://doi.org/10.1021/acsomega.3c05224

As such, Olsthoorn et al. conducted a valid prior study but did not compare the graph representations. Because molecular crystals are composed of organic molecules, molecular and crystal graph representations are possible (Figure 1). The complexity of the graph also affects prediction performance, and it is crucial to investigate which graph representation has the much better prediction accuracy. Furthermore, the difference between molecular and crystal graphs will allow quantitative evaluation of the influence of the molecular structure and intermolecular interactions on the target properties, which can provide insight into the physicochemical aspect.

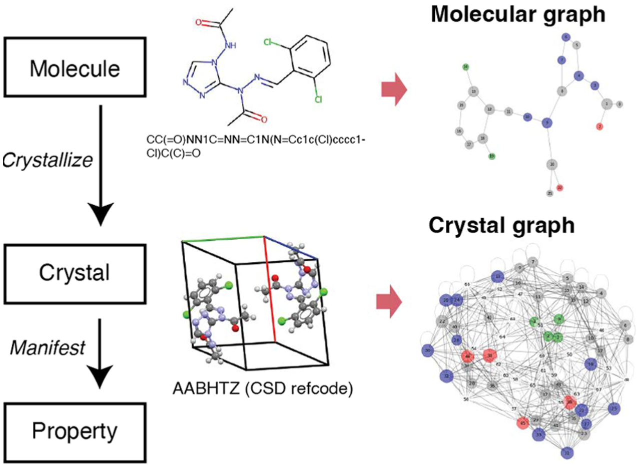


Figure 1. Structure−property relationship and graph representations of molecular crystals. Molecular crystals are formed by organic molecules and thus can be represented by both molecular and crystal graphs.

This study compares the prediction accuracy of molecular and crystal graphs in band gap prediction. Molecular and crystal graphs of varying complexities were used as input, and GNN models of multiple architectures were used as prediction functions. A comparison of regression results shows that crystal graphs give better prediction accuracy than molecular graphs and the contribution of molecular structure and intermolecular interactions is quantitatively evaluated. The novelty of this study lies in the comparison of molecular crystal representations and in the quantitative evaluation of the contribution of crystal structures to the band gap. Since this algorithm can be applied to other properties of molecular crystals, it is expected to contribute to the efficient screening and design of molecular crystals.

■ RESULTS AND DISCUSSION

Data Set and Structure Representations. A data set where the crystal structure and calculated electronic band gap corresponded was obtained from the Organic Materials Database (OMDB).[24] In the data set, there is no distinction between direct and indirect band gaps; that is, the value was defined as a distance between the minimum energy of the lowest conduction band and the maximum energy of the highest valence band independently. As the original data set did not contain simplified molecular input line entry system (SMILES) corresponding to each molecular structure, SMILES data were extracted from the COD database[29] for this study. Data that caused errors in processing SMILES conversion using rdkit package were excluded, yielding a data set of N = 10472, where SMILES, the crystal structure (as COD ID), and the band gap corresponded. The data set was split into train, validation, and test subset at the ratio of 0.8, 0.05, and 0.15, respectively. All subsets have a similar distribution of band gap (Figure 2).

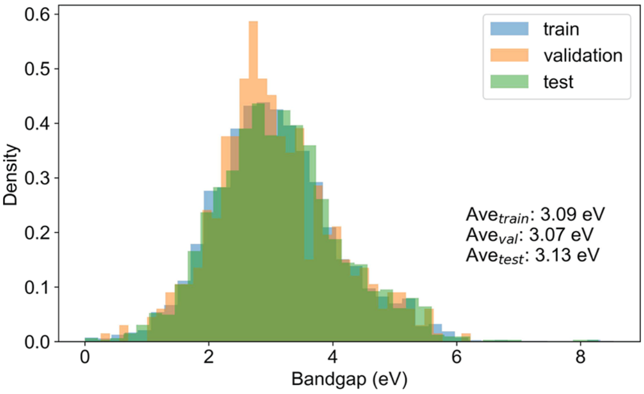
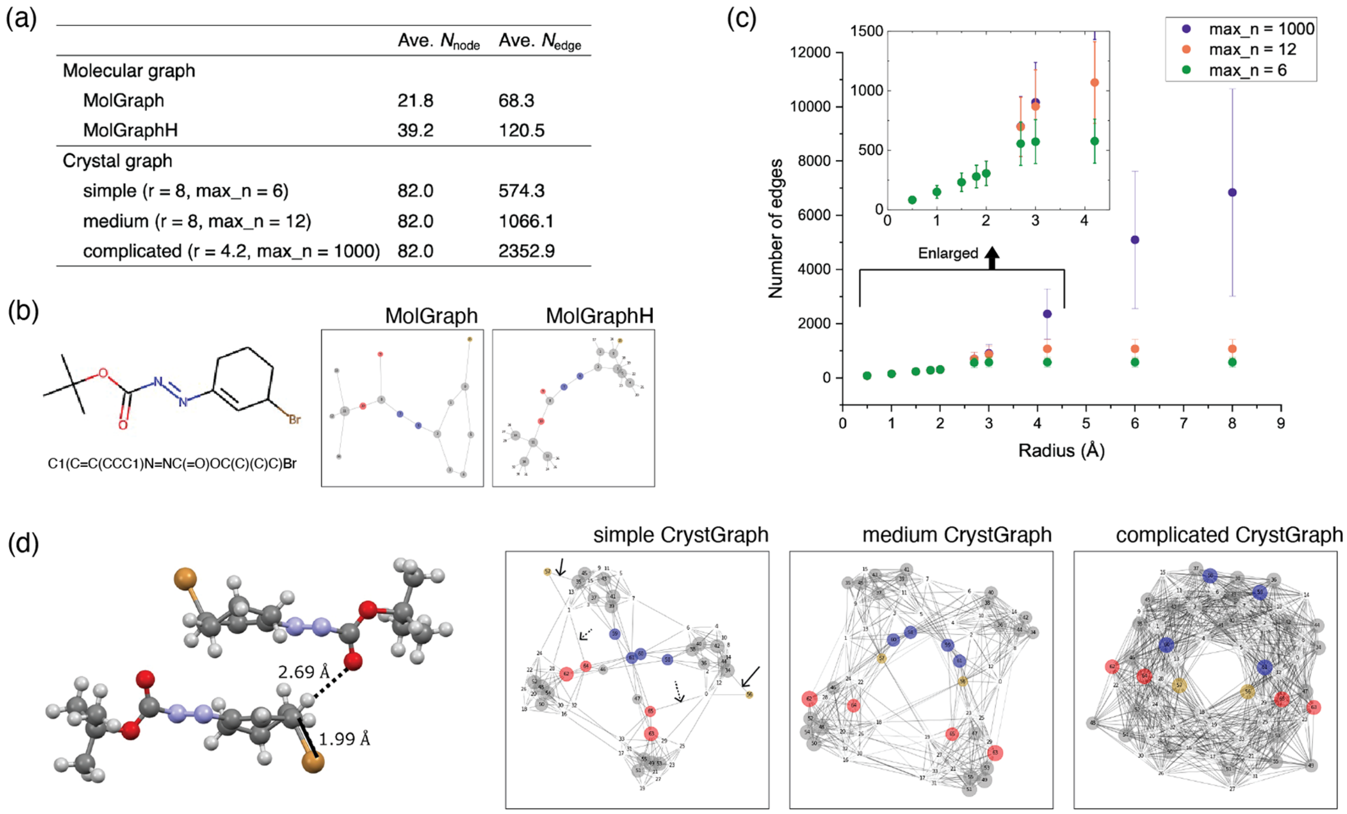


Figure 2. Data distribution of the band gap in the train, validation, and test subsets.

Molecular and crystal graphs are formed with different complexities. A SMILES-based molecular graph comprises nodes and edges that reflect atoms and chemical bonds (Table S1). Each node has a 121-dimensional representation that includes atomic information, such as the atom number and charge. Each edge has a 2-dimensional vector based on chemical bonds. As SMILES does not explicitly include hydrogen atoms, it is possible to determine whether they are added to molecular graphs. In the molecular graphs without hydrogen atoms (MolGraph), the number of nodes and edges averages 21.8 and 68.3, respectively (Figure 3a). These averages of molecular graphs with hydrogen atoms (MolGraphH) increased nearly twice (Figure 3a). Figure 3b exemplifies the MolGraph and MolGraphH of a molecule.

Crystal graphs were created based on atomic coordinates in the crystal structures. The main difference of the crystal graphs from molecular graphs is that edges are formed based on distances rather than chemical bonds. When edges are created in the crystal graphs, there are two parameters: radius and the maximum number of neighbors. The radius defines the maximum distance from one atom to the other atoms for sharing the edges. When the atom−atom distance is within the radius, an edge is allowed between these atoms. The maximum number of neighbors defines the maximum number of edges from a node, and the number of edges is limited to the nearest atoms, even when the atom−atom distance is within the radius. Thus, the number of edges in the crystal graphs depends on the radius (r) in the unit of Angstrom and the maximum number of neighbors (max_n). We considered three different crystal graphs: simple (r = 8 Å, max_n = 6), medium (r = 8 Å, max_n = 12), and complicated (r = 4.2 Å, max_n = 1000) graphs by changing these parameters (Figure 3a,c). Hereafter, they are called simple, medium, and complicated CrystGraphs. The average of the number of edges almost doubled from simple to medium CrystGraph and from medium to complicated CrystGraph (Figure 3a,c). This complexity graph could be achieved with other combinations of r and max_n, but we chose these combinations to vary the number of edges. The number of edges corresponds to the graph



	Ave. N_{node}	Ave. N_{edge}
Molecular graph		
MolGraph	21.8	68.3
MolGraphH	39.2	120.5
Crystal graph		
simple (r = 8, max_n = 6)	82.0	574.3
medium (r = 8, max_n = 12)	82.0	1066.1
complicated (r = 4.2, max_n = 1000)	82.0	2352.9

Figure 3. Molecular and crystal graphs with different complexities. (a) Comparison of the averages of the number of nodes and edges in different molecular and crystal graphs. (b) Example of molecular graphs without and with hydrogen atoms (MolGraph and MolGraphH, respectively). (c) Dependence of the average of the number of edges in crystal graphs by changing the radius (*r*) and maximum number of neighbors (*max_n*). The number of nodes does not change by these parameters. (d) Example of the crystal structure (COD ID = 4030612) and corresponding simple, medium, and complicated CrystGraphs.

Table 1. Regression Results of Band Gap Analysis Using Molecular and Crystal Graphs[a]

representation	GNN	R^2	MAE	RMSE
MolGraph	SchNet	0.732 (0.004)	0.401 (0.002)	0.527 (0.004)
	MEGNet	0.776 (0.006)	0.360 (0.003)	0.482 (0.007)
	CGCNN	0.754 (0.011)	0.375 (0.011)	0.505 (0.011)
MolGraphH	SchNet	0.693 (0.006)	0.427 (0.005)	0.564 (0.006)
	MEGNet	0.758 (0.014)	0.379 (0.013)	0.501 (0.015)
	CGCNN	0.750 (0.007)	0.385 (0.007)	0.509 (0.007)
simple CrystGraph	SchNet	0.862 (0.008)	0.279 (0.009)	0.378 (0.011)
	MEGNet	0.892 (0.006)	0.249 (0.014)	0.335 (0.010)
	CGCNN	0.858 (0.007)	0.283 (0.005)	0.383 (0.009)
medium CrystGraph	SchNet	0.833 (0.055)	0.308 (0.055)	0.412 (0.066)
	MEGNet	**0.895 (0.004)**	**0.240 (0.005)**	**0.329 (0.006)**
	CGCNN	0.856 (0.010)	0.280 (0.014)	0.386 (0.013)
complicated CrystGraph	SchNet	0.810 (0.084)	0.331 (0.072)	0.437 (0.094)
	MEGNet	0.879 (0.007)	0.259 (0.006)	0.354 (0.010)
	CGCNN	0.870 (0.004)	0.277 (0.004)	0.367 (0.006)

[a]Each value is the average, and the bracket represents standard deviation.

complexity, indicating the number of intermolecular interactions considered in graph convolutions. The crystal structure (COD number: 4030612), whose molecule is shown in Figure 3b, exemplifies the difference in graph complexity (Figure 3d). In the crystal structure, C−Br is the longest chemical bond (1.99 Å), and O···H is a short intermolecular interaction (2.69 Å) (Figure 3d). Simple CrystGraph captures the longest chemical bond and the short interaction by forming edges as indicated by solid and dotted arrows, while other longer intermolecular interactions are not considered. Medium CrystGraph captures more intermolecular interactions by a greater number of edges, and complicated CrystGraph contains much more. The difference in the graph complexity on the prediction performance is shown in the following section.

Regression on Band Gap. Molecular and crystal graphs with different complexities were input into the GNN models CGCNN, SchNet, and MEGNet.[6−8] The difference between

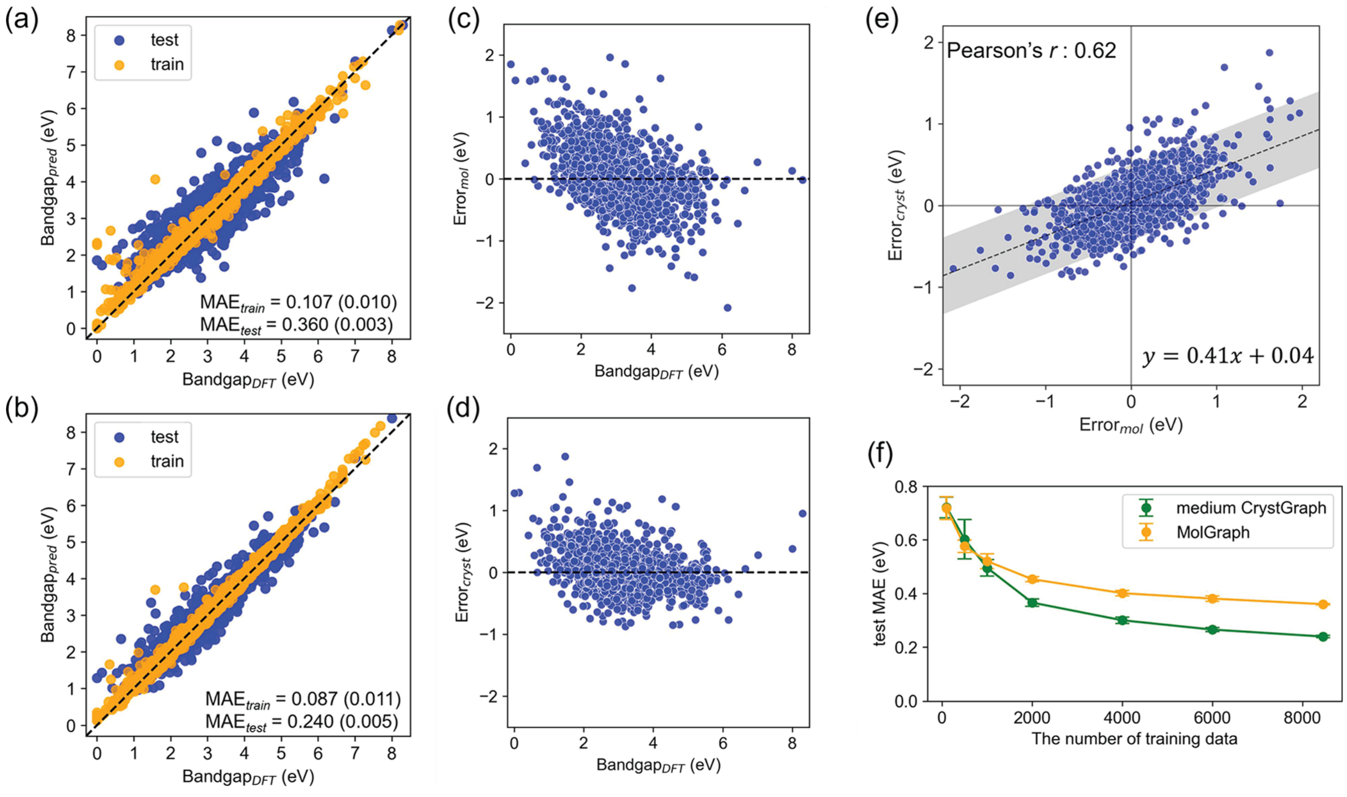


Figure 4. Predictive performance using molecular and crystal graphs. (a, b) Observed-predicted plots of training and test data set represented by (a) MolGraph and (b) medium CrystGraph. The dashed line is the reference line when predictions perfectly match with DFT data. (c, d) Error analysis of test data set represented by (c) MolGraph and (d) medium CrystGraph. Error was defined as predicted value minus observed value. (e) Error-error plot of MolGraph and medium CrystGraph. The dashed line is the linear regression. The area highlighted in gray is the 95% prediction interval. (f) Dependence of training size on test MAE.

these models relies on the graph convolution layer, which aggregates and combines node and edge features without changing the graph adjacency (Methods section). The node and edge features after convolution layers were converted to a vector by a readout operation, and the vector was then inputted to the fully connected layers to output a prediction value. The predictive performance was evaluated using the test metrics R^2, root-mean-square error (RMSE), and mean absolute error (MAE). The hyperparameters of all of the models were optimized with respect to each input representation (Table S2 and Figures S1−S15).

Table 1 presents the metrics for the test subset using different combinations of graphs and models. The MEGNet models performed better than SchNet and CGCNN in all combinations using molecular and crystal graphs (Table 1). Thus, the difference in graph inputs on test metrics is discussed based on the results of the MEGNet model.

The best representation was medium CrystGraph. Simple CrystGraph afforded nearly high performance with that of medium CrystGraph, and complicated CrystGraph was worse than those CrystGraphs (Table 1). Both molecular graphs, MolGraph and MolGraphH, afforded metrics that were worse than those of CrystGraphs. The better representation was MolGraph rather than MolGraphH. In both cases of molecular and crystal graphs, the most complicated graph did not afford the highest metrics, showing that there is a suitable complexity of representation. In the case of the average model, which represents the reference metrics when assuming there is no relationship between structure and band gap, MAE and RMSE were 0.794 and 1.018 eV, respectively. It is estimated that the combination of MolGraph and the MEGNet model reduced 0.63 eV in MAE by considering molecular structure and that the combination of medium CrystGraph and the MEGNet model further reduced 0.12 eV in MAE by considering intermolecular interactions. Thus, the error using medium CyrstGraph was 0.67 times that of MolGraph. It is reasonable that the intermolecular interactions are weak in many molecular crystals and thus have a moderate effect on the electronic orbitals by crystal packings, and the molecular structure alone can greatly reduce the prediction error. The combination of medium CrystGraph and MEGNet outperformed the ensemble model of SOAP and SchNet developed by Olsthoorn et al. (MAE = 0.388 eV, RMSE = 0.519 eV).[24] This result agreed with SchNet yielding worse metrics in our regression (Table 1). Although this may be due to the reduced number of data, the tuning of crystal graph hyperparameters and the utilization of the MEGNet model afforded better predictive performance. Then, we also performed the regression on the recent architecture, ALIGNN, using the same training data set and found that the MAE for the same test data set was 0.221 eV, which was slightly better than the MEGNet model. Based on this result, it is expected that the latest models such as ALIGNN and M3GNet would result in a smaller prediction error for molecular crystals. However, since it is difficult to input molecular graph used in the current work to these models, and since we want to compare the molecular graph with the crystal graph, we will use the MEGNet model in the following discussion.

Since MolGraph and medium CrystGraph afforded better test metrics as molecular and crystal representations, we

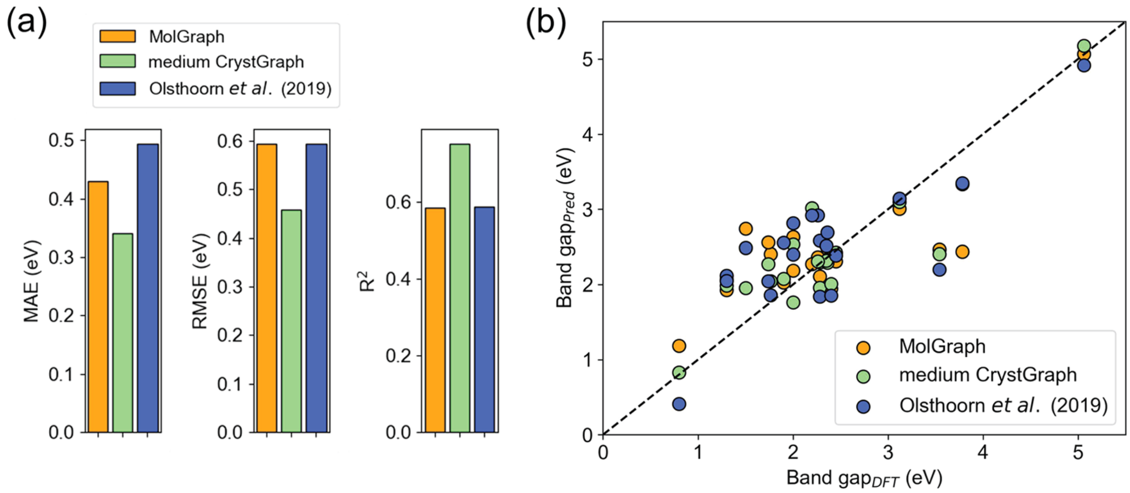


Figure 5. Generalization performance. (a) Comparison of the test metrics between MolGraph and medium CrystGraph developed by this work and the ensemble model developed in the literature. (b) Observed-predicted plot of the generalization data set.

performed error analysis of them where error was defined as predicted value minus observed value. First, the observed-predicted plot showed how predictions distributed (Figure 4a,b). The observed-predicted plot of medium CrystGraph presented a smaller variance from the reference line than MolGraph, which is consistent with better test metrics (Table 1). Both training and test MAEs of medium CrystGraph were lower than those of MolGraph, indicating proper representation learning of crystal structures. Even though, predictions of test data ranging from 1 to 5 eV, which was the main fraction of data set (Figure 2), have some variance from the reference line compared to the training data set. Error analysis of MolGraph showed the negative tendency of error (Figure 4c), meaning that MolGraph tends to overestimate for smaller band gaps and to underestimate for larger band gaps. This tendency should indicate there remains bias by missing information on intermolecular interactions. Medium CrystGraph resolved such a negative tendency (Figure 4d). This reason should be the capture of more information on CrystGraph. Crystal graph captures both intermolecular interactions and intramolecular interactions in the edge information, which may have reduced the prediction error. However, we cannot deny that other unidentified factors may influence the prediction.

We then visualized an error−error plot to find the error relationship between MolGraph and medium CrystGraph (Figure 4e). The error−error plot showed roughly linear tendency with a slope of 0.41 and Pearson's correlation coefficient of 0.62. This result means that medium CrystGraph afforded about 0.4 times smaller error than MolGraph with a mild correlation. This value is smaller than the value seen in the MAE because the MAE averages the absolute errors of all plots, whereas the error−error plot shows the relationship between the medium CrystGraph and MolGraph at the same data point. As a result of direct comparison, we can say that the medium CrystGraph has an error reduction of 0.4 times that of the MolGraph. However, since the correlation coefficient is not high and some predictions do not follow this trend, the MAE reduction effect is considered to have been 0.67 times. We did not identify the relationship between error distribution and structural features even though we checked the crystal structures manually of some outliers. If we find some relationship, we can speculate the strength and weakness of the crystal graph and possibly improve the convolution architecture of GNN, but we did not find so far.

The dependence of the training size on the test MAE was also compared between MolGraph and medium CrystGraph (Figure 4f). MolGraph converged faster than medium CrystGraph probably due to the simpler representation of graph. Medium CrystGraph reached a smaller MAE than MolGarph. This result showed the difference of convergence speed due to the difference of graph complexity and the sufficiency of training when sufficient training size was provided.

Generalization Ability and Large Screening. Subsequently, we confirmed the generalization ability outside the above data set. We collected 21 data from publications where SMILES, crystal structure, and calculated electronic band gap were reported (Table S3). The distribution of the number of constituent elements in newly collected data is similar to the initial data set for model training and testing (Figure S16). In addition, when crystal structures were represented by SOAP and then embedded by t-Distributed Stochastic Neighbor Embedding (t-SNE), the newly collected test data, of which there are only 21, is not widely distributed throughout the space, but it is found to be close to the initial data set (Figure S17). Using the trained MEGNet model, medium CrystGraph yielded better metrics (RMSE = 0.46, MAE = 0.34) than MolGraph (RMSE = 0.59, MAE = 0.43), confirming the better generalization ability of the combination of medium CrystGraph and the MEGNet model (Figure 5a). For further comparison, we calculated predicted values using the ensemble model of SchNet and SOAP developed by Olsthoom et al. on their web application.[23] The metrics of the predictions were comparable with those of MolGraph and worse than those of medium CrystGraph (Figure 5a). The observed-predicted plot showed how predictions are made for each data, and light green plots of medium CrystGraph were tented to be closer to the reference line (Figure 5b), confirming the better generalization. This generalization is consistent with the regression results, as explained in the previous section.

We also tested the prediction ability for polymorphic crystals using medium CrystGraph, whose compounds are not included in the training data set (Table S4). There have been 8 polymorphs of 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile, commonly known as ROY, and their electronic band gaps have been calculated to be 1.12−1.84 eV depending on crystal structures.[30] In another case, band gaps of two polymorphs of 9,10-bis((E)-2-(pyridin-4-yl)vinyl)-anthracene (BP4VA) have been reported to be 1.58 and 1.26 eV in the literature.[31] Using the trained model, medium CrystGraph of ROY afforded 0.40 eV MAE, and BP4PV resulted in 0.52 eV MAE (Figure 6). The result, a smaller error of ROY than BP4VA, may be attributed to the difference in a polymorphic manner. In the ROY crystal, molecular conformations differ from each polymorph due to the molecular flexibility, while the stacking manner of relatively rigid molecules differs in the polymorphs of BP4VA. Such polymorphic differences may have resulted in different MAEs. In addition, these values were worse than the previous generalization test, and this result suggests that medium Crystal may have a weakness in the distinction of polymorphs. Error plot also presented a negative tendency, suggesting insufficient learning. This result may originate from the deficiency of training data of polymorphic crystals because the original OMDB database does not contain polymorphic crystals. Such model extensions to the polymorphs should be tackled in the future.

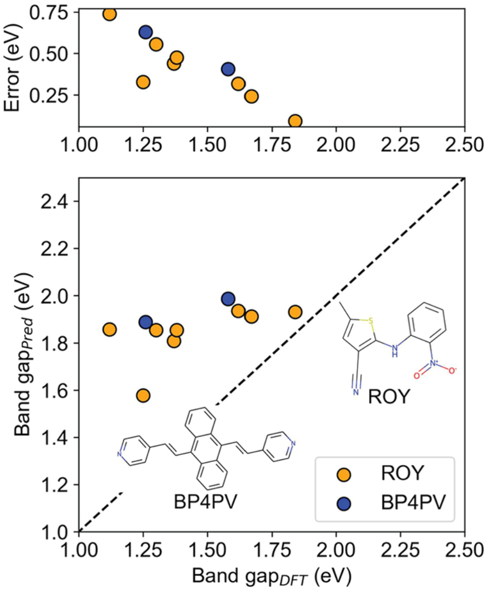


Figure 6. Band gap prediction of polymorphic crystals, ROY and BP4VA, using a medium CrystGraph.

Since medium CrystGraph afforded better generalization ability, we screened a large data set of crystal structures downloaded from the Cambridge Structural Database (CSD). It is important to find molecular crystals with high and low band gaps by screening for efficient material design. The number of downloaded crystal structures is 111936, identified with the unique CSD reference code. The distribution of the data set for screening was similar to that of the initial data set, after their structures encoded by SOAP descriptor were embedded in 2D plot (Figure S18). The band gap, predicted by medium CrystGraph on the trained MEGNet model, is distributed with a mean of 2.86 eV, a maximum of 7.39 eV, and a minimum of 0.13 eV, respectively (Figure 7a). This distribution is similar to the data distribution curated from the OMDB (Figure 2).

Some crystal structures were visualized to find structural features from the screening. The smallest predictions were found in compounds with sulfur atoms, as exemplified by the top-3 smallest predictions (Figure 7b). Two of the three structures consist of thiophene-based compounds, and the smallest prediction is consistent with that thiophene-based compounds such as tetrathiafulvalene (TTF) salt are known as conductive materials.[32] The largest predictions were found in the hydrocarbon and similar compounds (Figure 7b). Because the crystal of propane has the highest band gap (8.54 eV) in the original data set, the inferred structures reflected the trend of a high band gap. The inferred band gap in the large screening was almost consistent with the known structural features, and therefore, we successfully screened potential crystal structures with varying band gaps. All inferred band gaps were available at https://github.com/takuyhaa/OrgCrystGNN/.

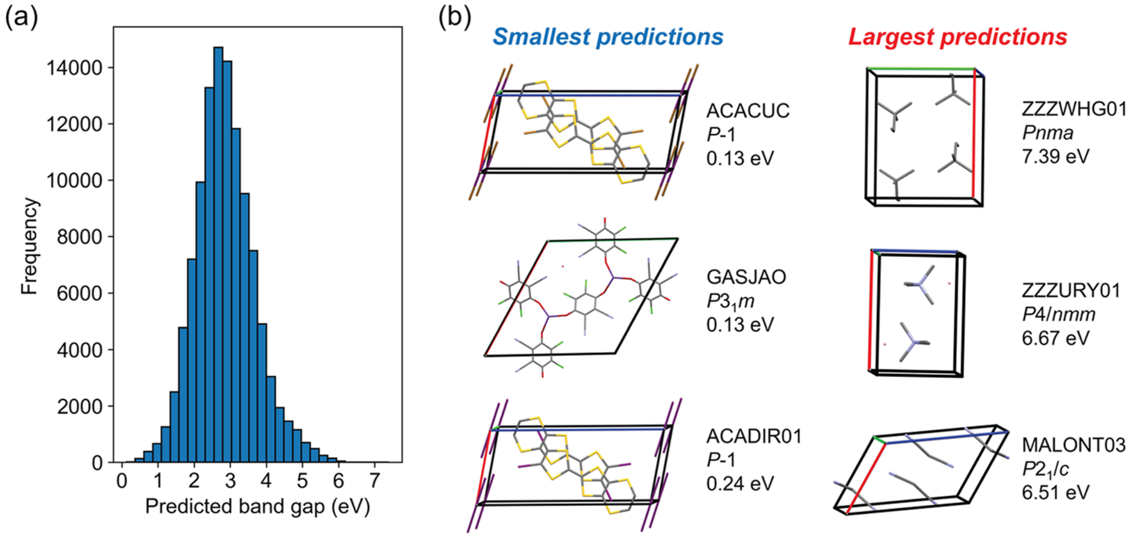


Figure 7. Comprehensive screening of band gap of crystal structures from CSD. (a) Histogram of band gap predicted by the trained model of medium CrystGraph. (b) Top-3 crystal structures with the smallest and largest predicted band gap. CSD reference code, space group, and predicted band gap are also shown.

■ CONCLUSIONS

In summary, this work compares graph representations of molecular crystals for the exemplified task of band gap prediction. The representation of medium CrystGraph afforded the best regression metrics compared with molecular graphs and other crystal graphs with different complexities. While this result is not surprising, it was beneficial to find that medium CrystGraph reduced error by about 0.4 times with a moderate correlation coefficient than MolGraph. A generalization test using manually collected data validated better performance using medium CrystGraph than MolGraph and the ensemble model developed in the literature. The band gap of large data set downloaded from CSD was screened, and the reasonable inference result was obtained, identifying potential crystals with lowest and highest band gaps. The novelty of this work should lie in clarifying the relative effect of crystal graph over molecule graph through the representation comparison. This workflow can be applied to other properties, potentially contributing to the efficient screening and design of functional molecular crystals.

■ METHODS

Data Collecting. The original data set was obtained from the Organic Materials Database (OMDB). The data set contains 12500 data, where the crystal structure and band gap corresponded. The crystal structures were assigned as the ID number of the Crystallography Open Database (COD). The DFT calculation has been performed using the projector augmented wave method[33,34] implemented in the Vienna Ab initio Simulation Package (VASP). The exchange-correlation function was approximated by the generalized gradient approximation (GGA) according to Perdew, Burke, and Ernzerhof.[35] The band gap was calculated after each crystal structure was geometrically optimized. Because this data set did not contain SMILES, we independently downloaded a list of SMILES with COD ID. The SMILES was added to the data set when the COD ID was matched. Additionally, the data set was modified to exclude data that arise execution errors in the conversion from SMILES to mol object using rdkit package. This modification finally afforded a data set of $N = 10472$, where SMILES, crystal structure, and band gap correspond. For the generalization test, we collected 21 data from published works, where SMILES and CCDC reference codes and band gaps were reported (Table S3). For large screening, we downloaded the cif files from Cambridge Structural Database (CSD) by the following conditions: only organic, calculated density larger than zero, temperature is not none, R factor less than 0.1, no disorder, and atmospheric pressure. The downloaded data contained various measurement temperatures, and thus, data were limited to data measured in the range of 273−313 K. This curation resulted in data size of 111936 corresponded to unique CSD reference codes.

Molecular and Crystal Representations. For molecular graphs (MolGraph and MolGraphH), SMILES was converted to mol object using the rdkit package, and node and edge features were stored as graph data using PyTorch and PyTorch-Geometric. For crystal graphs (simple, medium, and complicated CrystGraphs), the atomic coordinates in the crystal structures were obtained from cif or json file and then converted to graph data using PyTorch and PyTorch-Geometric.

Graph Neural Networks. The GNN models (CGCNN, SchNet, and MEGNet) were originally developed in independent studies.[6−8] These GNN models commonly consist of convolution layers, readout operation, and a fully connected neural network. The difference in GNN models relies solely on convolutional operations that update node and edge features without changing graph adjacency.

In the SchNet,[6] the convolution operation is

$$\mathbf{x}_i' = \sum_{j \in N(i)} \mathbf{x}_j \odot h_\Theta(\exp(-\gamma(\mathbf{d}_{j,i} - \mu)))$$

where $\mathbf{x}_i'$ is the aggregated feature added to node feature $\mathbf{x}_i$, $\mathbf{x}_j$ is the adjacency node feature of node i, h_Θ denotes a multiple linear perceptron (MLP), $\mathbf{d}_{j,i}$ denotes the interatomic distances between atoms, μ is values for Gaussian expansion, and γ is a coefficient. In this work, we set $0 \le \mu \le 1$ for equally spaced sampling ($n = 50$) and $\gamma = -12.5$.

In the CGCNN model,[7] the convolutional operation is

$$\mathbf{x}_i' = \mathbf{x}_i + \sum_{j \in N(i)} \sigma(\mathbf{z}_{i,j}\mathbf{W}_f + \mathbf{b}_f) \odot g(\mathbf{z}_{i,j}\mathbf{W}_s + \mathbf{b}_s)$$

where $\mathbf{z}_{i,j} = [\mathbf{x}_i, \mathbf{x}_j, \mathbf{e}_{i,j}]$ denotes the concatenation of central node features, neighboring node features and edge features. Here, $\mathbf{e}_{i,j} = \exp(-\gamma(\mathbf{d}_{j,i} - \mu))$. In addition, σ and g denote the sigmoid and softplus functions, respectively. $\mathbf{W}$ and $\mathbf{b}$ represent the weight and bias matrices, respectively.

In the MEGNet model,[8] the convolution operation is

$$\mathbf{e}_{i,j}' = h_{\Theta_e}(\mathbf{x}_i \oplus \mathbf{x}_j \oplus \mathbf{e}_{i,j})$$

$$\mathbf{x}_i' = h_{\Theta_v}\left(\left(\frac{1}{N(i)} \sum_{j \in N(i)} \mathbf{e}_{i,j}\right) \oplus \mathbf{x}_i\right)$$

where h_{Θ_e} and h_{Θ_v} are the edge and node update functions of the MLP, respectively. Two dense layers were added at the beginning of each convolution layer to preprocess the inputs. The update operation is performed in the order of the edges, nodes, and global attributes. These GNN models have the following hyperparameters: the dimension of the dense layer for preprocessing in convolution, the dimension of the update function (MLP), the dimension of the MLP after the readout operation, the number of convolution layers, the number of dense layers after the readout, the pooling operation, and the learning rate. The batch size and number of epochs were fixed to 64 and 200, respectively. Hyperparameter optimization was performed using Optuna in a manually defined search space (Table S2). GNN models were trained to minimize L1 loss on a Windows computer equipped with GPUs (NVIDIA RTX A6000). The codes were written based on PyTorch and PyTorch-Geometric libraries. Codes are available at https://github.com/takuyhaa/OrgCrystGNN/.

The model training was repeated three times, and the test metrics were calculated by averaging three results. For the dependency of training size on test metrics, three different training data sets with a fixed data size were created by different random state and three trainings were performed on a training data set. In all cases, the same test data set was used for evaluation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05224.

Hyperparameters, data set for generalization test, data distribution (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Takuya Taniguchi − *Center for Data Science, Waseda University, Tokyo 169-8050, Japan;* orcid.org/0000-0002-7885-2962; Email: takuya.taniguchi@aoni.waseda.jp

Authors

Mayuko Hosokawa − *Department of Advanced Science and Engineering, Graduate School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan*

Toru Asahi − *Department of Advanced Science and Engineering, Graduate School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05224

Author Contributions

T.T. contributed to conceptualization, investigation, software, formal analysis, visualization, writing − original draft, writing − review and editing, funding acquisition, project administration, supervision. M.H. contributed to data curation. T.A. contributed to supervision for M.H.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partly executed under the cooperation of organization between Waseda University and ENEOS Corporation. This work was financially supported by the JSPS Grant-in-Aid (22K14747) and Waseda University Grant for Special Research Projects (2020C-530, 2021C-404, and 2022C-313).

■ REFERENCES

(1) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **2017**, *3*, 54.

(2) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(3) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.

(4) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(5) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **2021**, *121*, 9759−9815.

(6) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. Schnet−a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, No. 241722.

(7) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, No. 145301.

(8) 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.

(9) Jiang, Y.; Yang, Z.; Guo, J.; Li, H.; Liu, Y.; Guo, Y.; Li, M.; Pu, X. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **2021**, *12*, No. 5950.

(10) Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **2020**, *4*, No. 063801.

(11) 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.

(12) Cheng, J.; Zhang, C.; Dong, L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater.* **2021**, *2*, 92.

(13) Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185.

(14) Pandey, S.; Qu, J.; Stevanović, V.; John, P. S.; Gorai, P. Predicting energy and stability of known and hypothetical crystals using graph neural network. *Patterns* **2021**, *2*, No. 100361.

(15) Frey, N. C.; Akinwande, D.; Jariwala, D.; Shenoy, V. B. Machine learning-enabled design of point defects in 2d materials for quantum and neuromorphic information processing. *ACS Nano* **2020**, *14*, 13406−13417.

(16) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **2021**, *7*, 84.

(17) Jalali, M.; Tsotsalas, M.; Wöll, C. MOFSocialNet: Exploiting metal-organic framework relationships via social network analysis. *Nanomaterials* **2022**, *12*, 704.

(18) Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2022**, *2*, 718−728.

(19) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity statistical machine learning for molecular crystal structure prediction. *J. Phys. Chem. A* **2020**, *124*, 8065−8078.

(20) Han, Y.; Ali, I.; Wang, Z.; Cai, J.; Wu, S.; Tang, J.; Zhang, L.; Ren, J.; Xiao, R.; Lu, Q.; Hang, L.; et al. Machine learning accelerates quantum mechanics predictions of molecular crystals. *Phys. Rep.* **2021**, *934*, 1−71.

(21) Ishizaki, K.; Sugimoto, R.; Hagiwara, Y.; Koshima, H.; Taniguchi, T.; Asahi, T. Actuation performance of a photo-bending crystal modeled by machine learning-based regression. *CrystEngComm* **2021**, *23*, 5839−5847.

(22) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. *Acta Crystallogr., Sect. B* **2016**, *72*, 171−179.

(23) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Bail, A. L. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **2012**, *40*, D420−D427.

(24) Olsthoorn, B.; Geilhufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band gap prediction for large organic crystal structures with machine learning. *Adv. Quantum Technol.* **2019**, *2*, No. 1900023.

(25) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.

(26) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12*, 4536−4546.

(27) Takagi, D.; Ishizaki, K.; Asahi, T.; Taniguchi, T. Molecular screening for solid−solid phase transition by machine learning. *Digital Discovery* **2023**, *2*, 1126.

(28) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine learning for the structure−energy−property landscapes of molecular crystals. *Chem. Sci.* **2018**, *9*, 1289−1300.

(29) Quirós, M.; Gražulis, S.; Girdzijauskaitė, S.; Merkys, A.; Vaitkus, A. Using SMILES strings for the description of chemical

connectivity in the Crystallography Open Database. *J. Cheminf.* **2018**, *10*, 23.

(30) Feng, X.; Becke, A. D.; Johnson, E. R. Theoretical investigation of polymorph-and coformer-dependent photoluminescence in molecular crystals. *CrystEngComm* **2021**, *23*, 4264−4271.

(31) Aziz, A.; Sidat, A.; Talati, P.; Crespo-Otero, R. Understanding the solid state luminescence and piezochromic properties in polymorphs of an anthracene derivative. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2832−2842.

(32) Odom, S. A.; Caruso, M. M.; Finke, A. D.; Prokup, A. M.; Ritchey, J. A.; Leonard, J. H.; White, S. R.; Sottos, N. R.; Moore, J. S. Restoration of conductivity with TTF-TCNQ charge-transfer salts. *Adv. Funct. Mater.* **2010**, *20*, 1721−1727.

(33) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953.

(34) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758.

(35) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.