Materials Informatics of Molecular Crystals Compared by Molecular and Crystal Representations: The Case of Band Gap Prediction

Takuya Taniguchi,*[a] Mayuko Hosokawa,[b] and Toru Asahi[b]

Abstract: In materials informatics, the representation of material structures is fundamentally important to obtain better prediction results. Molecular crystals can be represented by both molecular and crystal representations, but there has been no examination to determine which representation is the most effective for the materials informatics of molecular crystals. In this work, different representations for molecular crystals were compared in an exemplified task of band gap prediction. We demonstrated that the predictive ability using molecular graph outperformed those of molecular fingerprints and crystal graphs. This result motivated the screening of molecular big data from PubChem, and the inference suggested candidate molecules of organic semiconductors for photovoltaics and luminescence. The novelty of this work relies on the representation comparison of molecular crystals and the finding that molecular graph works better even though the property prediction of crystalline materials. This finding will enable to machine-learning-aided screening and design of functional molecular crystals.

Introduction

Materials informatics (MI) is a hot research topic in both academia and industries.[1,2,3] The application of MI is becoming widespread, and MI tasks are mainly divided into regression and classification. Regression is the task of predicting the property values of unseen materials based on the unknown relationship between known materials and properties, with the aim of finding a better prediction function \( f \) in the equation \( y = f(x) \), where \( x \) is the structural information of a material and \( y \) is the target property. Classification is the task of predicting class(es): for example, the crystal system, the presence of a specific chemical event, and whether two compounds interact. The aim of classification is also to find better functions for border lines to separate the data.

In most cases of MI, a set of structural information and target properties is required as a prior knowledge for model training. While a target variable can be defined relatively straightforward depending on the purposes, structural information can be represented as descriptors in various ways. Many types of descriptors have been developed, and the choice of descriptor(s) influences the predictive performance of a task to be solved. Therefore, it is crucial to determine how a material structure is treated as a descriptor. For molecules and polymers, conventional representations are fingerprint vectors, such as Avalon and extended connectivity fingerprints (ECFP).[4-8] These fingerprints can be obtained via string representations using a simplified molecular input line-entry system (SMILES). Graph representation is another typical format for molecules, and can be processed by graph neural networks (GNN) for regression and classification tasks.[9-12] In the case of other materials, such as inorganic crystals, a classical representation is the ratio and atom types of composites, and recent works have succeeded in using crystal graphs for inorganic solids.[13-17] Graph-based approaches have attracted increasing attention in material research, and have been applied not only to molecules and bulk inorganics, but also metal-organic frameworks, two-dimensional materials, and further material simulations.[18-20]

In contrast, MI research on organic molecular crystals has progressed to a lesser extent than polymers and inorganic crystals. This could be due to the shortage of databases containing molecular crystal structures and property data. For example, the Cambridge Structural Database (CSD) and the Crystallography Open Database (COD) are informative structural platforms for molecular crystals,[21,22] but do not include information on physicochemical properties. Thus, the construction of structure-property datasets requires human effort to collect property data from published works, experiments, and quantum chemical calculations. A recent study has constructed a large database \((N = 12,500)\) of the band gap, where the structures of molecular crystals correspond to the band gap obtained by quantum chemical calculations.[23] This dataset can be utilized in MI research to screen potential organic semiconductors for photovoltaics and luminescent materials.

Once such a database becomes available, the unique character of molecular crystals raises the question of how they should be represented. Molecular crystals consist of organic molecules, and thus, both molecular and crystal representations are possible (Figure 1). However, such descriptor comparisons have not been examined, although some studies on molecular crystals have utilized machine learning techniques for material design and experimental process.[9-13]
RESEARCH ARTICLE

Figure 1. Representations of molecular and crystal structures for machine learning. Molecular crystals are formed by organic molecules, and thus can be represented by both molecular and crystal representations. This work considers fingerprints and molecular graphs for molecular representation, and crystal graphs for crystal representation.

From a physicochemical point of view, solid-state properties such as the band gap correspond to the crystal structure, while molecular representations may provide sufficient predictive performance in the MI task. If molecular representations result in better performance than crystal representations, molecular information can be used to screen solid-state properties and suggest a wider range of candidates than has been possible thus far.

This work describes bandgap prediction, achieved by comparing molecular fingerprints, molecular graphs, and crystal graphs. Different types of fingerprints and molecular and crystal graphs with different complexities are considered as inputs for prediction, employing a neural network (NN) for fingerprints and GNN models for graph inputs, respectively. A comparison of the regression results revealed that a molecular graph representation performed best, including against crystal graphs. This result enabled screening of molecular big data from PubChem ($N > 6.8$ million) for band gap exploration, and candidate molecules for organic semiconductors have been proposed. As a suitable representation should depend on the task, it is expected that this workflow will contribute to the efficient screening and design of molecular crystals.

Results and Discussion

Dataset and Structure representations

A dataset, where the crystal structure and calculated band gap corresponded, was obtained from the Organic Materials Database (OMDB). As the original dataset did not contain SMILES corresponding to each molecular crystal, SMILES data was extracted and added to the dataset for this study. Then, multicomponent molecules were excluded for a simpler comparison of molecular and crystal representations. The addition and cleaning of the original dataset afforded a dataset of smaller size ($N = 8830$) with almost the same distribution (Figure 2a,b). In the modified dataset, SMILES, the crystal structure (as COD ID), and the band gap are corresponded.

Figure 2. Data distribution of band gap. (a) Distributions of datasets used in this work and the original database, OMDB. (b) Statistics of the datasets.

To represent the molecular structures, SMILES was converted to fingerprints and molecular graphs. The fingerprints are ECFP2, Avalon, ErG, MACCSKeys, and Estate; each fingerprint is a vector representation based on molecular substructures. The dimensions of these fingerprints differed depending on the algorithms or predefined substructures (Table S1). A molecular graph is composed of nodes and edges that reflect atoms and chemical bonds, respectively (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 include hydrogen atoms explicitly, it is possible to determine whether hydrogen atoms are added to molecular graphs. In the molecular graphs without hydrogen atoms (MolGraph), the averages of the number of nodes and edges are 19.3 and 41.9, respectively (Figure 3a). These averages of molecular graphs with hydrogen atoms (MolGraphH) increased nearly twice (Figure 3a). Figure 3b shows MolGraph and MolGraphH of a molecule.

To represent the crystal structures, the atomic coordinates were converted into crystal graphs. In the crystal graph, nodes also reflect atoms, while edges represent spatial distances...
expanded on a Gaussian basis, reflecting the distance between atoms in a crystal structure (Table S1).

### Table 1

<table>
<thead>
<tr>
<th>Crystal graph</th>
<th>Ave. N nodes</th>
<th>Ave. N edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 8, \max_n = 6 )</td>
<td>82.3</td>
<td>576.4</td>
</tr>
<tr>
<td>( r = 8, \max_n = 12 )</td>
<td>82.3</td>
<td>1070.5</td>
</tr>
<tr>
<td>( r = 4.2, \max_n = 1000 )</td>
<td>82.3</td>
<td>2357.0</td>
</tr>
<tr>
<td>Molecular graph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>without H atoms</td>
<td>19.3</td>
<td>41.9</td>
</tr>
<tr>
<td>with H atoms</td>
<td>34.7</td>
<td>72.6</td>
</tr>
</tbody>
</table>

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) An 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 radius \( r \) and maximum number of neighbors \( \max_n \). The number of nodes does not change by these parameters. (d) An example of crystal structure (COD ID = 4030612) and corresponding simple, medium, and complicated CrystGraphs.

The main difference between the crystal graph and the molecular graph is that edges are formed, based on distances rather than chemical bonds. To form edges in the crystal graphs, there are two parameters: radius \( r \) and maximum number of neighbors \( \max_n \). 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 formed between these atoms. The maximum number of neighbors define the maximum number of adjacent nodes, and edge sharing 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 \( r \) and the maximum number of neighbors \( \max_n \). We considered three different crystal graphs: simple \((r = 8, \max_n = 6)\), medium \((r = 8, \max_n = 12)\), and complicated \((r = 4.2, \max_n = 1000)\) 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 contains much more. The difference in the graph complexity on the prediction performance is shown in the following section.

#### Regression on bandgap

Molecular and crystal graphs with different complexities were input into the GNN models CGCNN, SchNet, and MEGNet. The difference between these models relies on the graph convolution layer, which aggregates and combines node and/or edge features without changing the graph adjacency (see Supporting Information). 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 NN to output a prediction value (Figure S1). In the case of fingerprints, the fingerprint vector is directly inputted to the NN. The predictive performance was evaluated using the test metrics \( R^2 \), root mean squared error (RMSE), and mean absolute error (MAE). The hyperparameters of all the models were optimized with respect to each input representation (Tables S2 and S3).

Table 1 presents the regression results for different combinations of inputs and models for a training size of 6000. Among the fingerprints, Avalon yielded better test metrics than...
other fingerprints. Although random forest was also used as a non-deep learning model for comparison, the metrics were lower than those of Avalon using the NN model (Table S4). In the case of graph inputs, the MEGNet models performed better than SchNet in all combinations using molecular and crystal graphs (Table 1). Although CGCNN for MolGraph and MolGraphH worked well, comparable to MEGNet, the results for CrystGraphs were not sufficient metrics. Thus, the difference in graph inputs on test metrics is discussed, based on the results using the MEGNet model, which afforded better metrics for molecular and crystal graphs.

MolGraph outperformed fingerprints and crystal graphs (Table 1). MolGraphH exhibited a performance similar to that of MolGraph. Medium CrystGraph afforded better test metrics than simple and complicated CrystGraphs and fingerprints, but did not achieve the performance of MolGraph and MolGraphH. When listing the three representations in the better order, MolGraph, MolGraphH, and medium CrystGraph are listed.

The dependence of the training size on the test metrics was measured to confirm the efficiency and appropriateness of the training. The comparison of all fingerprints with the NN model clarified the better improvement in MAE using Avalon (Figure 4a). Avalon afforded a faster improvement of test metrics than MolGraphs and CrystGraphs with MEGNet models (Figure 4c,e), suggesting that graph representations require a larger amount of data for sufficient training, in comparison to fingerprints. Among the graph representations, MolGraph and MolGraphH converged at almost the same speed and slightly faster than simple and medium CrystGraphs. Complicated CrystGraph afforded slower improvement in test metrics than other inputs, suggesting that the amount of data may not be sufficient for complicated CrystGraph. The observed-predicted plots when using Avalon, MolGraph, and medium CrystGraph clarified the higher predictive performance of MolGraph (Figure 4b,d,f). The dependences of the training size using other models (RF for fingerprints, CGCNN and SchNet for graphs) were also investigated, showing similar trends in improvement efficiency (Figures S2-4).
interactions, but the magnitude of the effect may be smaller than that of the molecular structure itself. In addition, CrystGraph contains spatial distance as edge features, whereas MolGraph contains chemical bond information. Such feature differences between MolGraph and CrystGraph might affect the predictive performance. Here, we mention again that the necessity of representation comparison comes from the character of molecular crystals, and this task is different with regression tasks about HOMO-LUMO energy difference of single molecules, tackled in several works using QM9 database.[29,30]

Figure 4. Dependence of the training size on test metrics. (a) Regression result by the combinations of fingerprints with the NN, and (b) the observed-predicted plot using Avalon at \( N_{\text{train}} = 6000 \). (c) Regression result using MolGraph and MolGraphH. (d) Observed-predicted plot using MolGraph at \( N_{\text{train}} = 6000 \). (e) Regression result using simple, medium, and complicated CrystGraph. (f) Observed-predicted plot using medium CrystGraph at \( N_{\text{train}} = 6000 \). For graph inputs, MEGNet model was used.

**Generalization ability and Molecular screening**

Subsequently, we validated the generalization ability outside the dataset. We manually collected 19 data from publications, where SMILES, crystal structure (as CCDC number), and band gap were reported (Table S5). MolGraph and medium CrystGraph were constructed from SMILES and crystal structures, respectively, and their inference results using trained MEGNet models were compared. MolGraph yielded better metrics (RMSE = 0.80, MAE = 0.66) than medium CrystGraph (RMSE = 1.00, MAE = 0.73). The comparison of the observed-predicted plot displayed a better prediction using MolGraph (Figure 5a), suggesting a higher generalization ability of MolGraph for predicting the band gap than CrystGraph.

The explainability of the GNN model is important for human understanding of how inferences are made. We used gradient-weighted class activation mapping (Grad-CAM), which was originally implemented for classification tasks[31] and later used for regression tasks of chemistry.[32,33] Grad-CAM reveals which nodes of the graph have positive or negative effects on the prediction. We chose an exemplified compound for comparison of molecular and crystal graphs (Figure 5b,c), based on the reason that a CrystGraph of a crystal structure with \( Z = 1 \) is relatively easy to compare with MolGraph. The MolGraph afforded a smaller error (\( y_{\text{obs}} = 1.77 \text{ eV} \) and \( y_{\text{pred}} = 1.74 \text{ eV} \)), and both positive and negative effects on the prediction were observed (Figure 5b). On the other hand, the medium CrystGraph resulted in larger error (\( y_{\text{obs}} = 1.77 \text{ eV} \) and \( y_{\text{pred}} = 3.29 \text{ eV} \)), and a positive effect was observed (Figure 5c). This positive single effect may lead to an overestimation of the band gap using medium CrystGraph, as seen in other crystals in the generalization test (Figure 5a).

Figure 5. Generalization performance and model interpretation. (a) The observed-predicted plot of newly collected data. MolGraph and medium CrystGraph were used as representations for molecular and crystal structures, and trained MEGNet models were used. Data indicated by arrows were depicted in panel b and c. (b) An exemplified molecular structure and corresponding MolGraph colored by weights of Grad-CAM. (b) The crystal structure of the same compound \( Z = 1 \) and corresponding medium CrystGraph colored by
weights of Grad-CAM. In the crystal structure, molecules stack via π-π interactions. Color bars indicate relative magnitude of positive and negative effects.

The finding that MolGraph affords the best predictive ability is useful because it means that the band gap of organic solids can be screened using molecules without prior knowledge of their crystal structures. 6.8 million single-component molecules as SMILES were collected from the PubChem database, and then band gap inference using MolGraph and the trained MEGNet model was performed for screening potential organic semiconductors. The distribution of the predicted values was similar to the normal distribution with an average of 3.3 eV (Figure 6a). From these inference results, some categories were selected to exemplify the similarity and diversity of molecules for potential semiconductors.

In the category of higher band gaps ($E_g > 7$ eV), it was found that there were many hydrocarbon molecules (Figure 6b). A total of 137 molecules were identified in this category. Because the crystal of propane (CH$_3$-CH$_2$-CH$_3$) has the highest band gap (8.2 eV) in the original dataset, the inferred molecules reflected the trend of a high band gap. The proposed hydrocarbons were not included in the original dataset, and most of their band gaps in the solid state have not been reported in the literature. Therefore, we successfully screened potential molecules for high band gap.

The second category ($1.7 < E_g < 1.9$ eV) was selected as a potential material for organic photovoltaics (Figure 6c). A recent paper on organic photovoltaics showed that many n- and p-type organic semiconductors possess this band gap range.$^{[34]}$ Our result show that this $E_g$ range contains approximately 135,000 molecules, many of which are π-conjugated (Figure 6c). This molecular feature is similar to the molecular structures used in organic photovoltaics.$^{[33]}$ Candidates for photovoltaic materials were screened based on band gap inference, and material selection will be possible when other requirements are further considered.

The third category ($0.5 < E_g < 1.4$ eV) shows potential molecules for near infrared (IR) luminescence (Figure 6d). This $E_g$ range corresponds to the near IR spectrum (800 – 2500 nm), and the luminescence of the near IR can be used for telecommunication and internal imaging. Band gap inference presented 40853 molecules in this range. There were many π-conjugated molecules as in the second category ($1.7 < E_g < 1.9$ eV), and the ratio of small molecules was higher than the second category. In the final category, molecules with smaller band gaps ($E_g < 0.5$ eV) may be potentially conductive matter, and this category contains 49 molecules, many of which contain sulfur atoms (Figure 6e).

**Conclusion**

In summary, this work compared various representations of molecular crystals for the exemplified task of band gap prediction. The representation of MolGraph afforded the best regression metrics, compared with molecular fingerprints and crystal graphs with different complexities. A generalization test using manually collected data validated better performance using MolGraph than medium CrystGraph. While this result may seem counterintuitive owing to the band gap being a solid-state property, the band gap should strongly depend on the molecular structure itself, thus enabling better predictive ability using MolGraph. This finding led to band gap inference using molecular big data from PubChem, and candidate molecules for organic semiconductors (as exemplified by the categories of high band gap, photovoltaics, and luminescent materials) were identified. This workflow can be applied to other properties, contributing to the efficient screening and design of functional molecular crystals.

**Acknowledgements**
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). This work was also partly supported by the Cabinet Office, Government of Japan, Cross-ministerial Moonshot Agriculture, Forestry and Fisheries Research and Development Program “Technologies for Smart Bio-industry and Agriculture” (BRAIN).

Conflict of interest

The authors declare no conflict of interest.

Data Availability Statement


Keywords: materials informatics • molecular crystals • graph neural network • deep learning • band gap


We compared molecular and crystal representations for the prediction task of band gap of molecular crystals, and found that molecular graph works better even though the prediction of crystal property. This finding enabled band gap inference using molecular data without knowing their crystal structures, suggesting candidate molecules for organic semiconductors.