Crysta1 Graph Neural Networks for Data Mining in Materials Science

Takenori Yamamoto*

RIMCS LLC, Yokohama, Japan

May 27, 2019

Abstract

Machine learning methods have been employed for materials prediction in various ways. It has recently been proposed that a crystalline material is represented by a multigraph called a crystal graph. Convolutional neural networks adapted to those graphs have successfully predicted bulk properties of materials with the use of equilibrium bond distances as spatial information. An investigation into graph neural networks for small molecules has recently shown that the no distance model performs almost as well as the distance model. This paper proposes crystal graph neural networks (CGNNs) that use no bond distances, and introduces a scale-invariant graph coordinator that makes up crystal graphs for the CGNN models to be trained on the dataset based on a theoretical materials database. The CGNN models predict the bulk properties such as formation energy, unit cell volume, band gap, and total magnetization for every test materials database. The CGNN models to be trained on the dataset based on a theoretical materials database. The CGNN models predict the bulk properties such as formation energy, unit cell volume, band gap, and total magnetization for every test-materialing material, and the average errors are less than the corresponding ones of the database. The predicted band gaps and total magnetizations are used for the metallic and non-magnet-magnet binary classifications, which result in success. This paper presents discussions about high-throughput screening of candidate materials with the use of the predicted formation energies, and also about the future progress of materials data mining on the basis of the CGNN architectures.

1 Introduction

The structure of a crystalline material \( S = (A, X, Z) \), and can approximately be calculated using the density functional theory (DFT).\(^1\,^2\) The approximate energy is hereafter expressed as \( E_{\text{tot}}(S) \). We can obtain the equilibrium structure by minimizing the total energy;

\[
\mathcal{S} = \arg \min_S E_{\text{tot}}(S) \text{ subject to } Z(S) \text{ fixed.} \tag{1}
\]

There are multiple minima, and \( E_{\text{tot}}(\mathcal{S}) \) is the equilibrium energy for a minimum. The DFT calculation also gives bulk properties such as the formation energy \( H_f \), the unit cell volume \( V_c \), the band gap \( E_g \), and the total magnetization \( M_t \). Note that all these are values at a temperature of 0 K and a pressure of 0 Pa.

There are theoretical materials databases containing relaxed structures of inorganic crystalline materials and their properties. The Materials Project (MP) database (V2018.12) holds 86k entries mostly related to experimentally observed structures.\(^3\,^4\) The Open Quantum Materials Database (OQMD v1.2) contains 563k entries related to experimental and hypothetical compounds.\(^5\,^6\)

We can train machine learning models that predict material properties in a supervised manner based on a materials database. The crystal graph convolutional neural networks (CGCNNs) were recently used with the MP database for this purpose.\(^7\) The crystal graph used in the CGCNN is a multigraph composed of the set of nodes (\( i.e., \) atoms) and the set of labeled edges (\( i.e., \) bonds). The edge label \( k \) is used as an identifier to distinguish between edges sharing both the endpoints owing to periodic boundary conditions. Every edge has distance features of the relevant bond. The CGCNN concatenates the node hidden states \( v_i \) and \( v_j \) and the distance feature vector \( u_{ij}^k \) for the edge \( e_{ij}^k \) as

\[
s_{ij}^k = v_i \oplus v_j \oplus u_{ij}^k, \tag{2}
\]

and uses them in the gated convolutions

\[
\sum_{j,k} \sigma(s_{ij}^k W_f + b_f) \odot g(s_{ij}^k W_s + b_s), \tag{3}
\]

where \( W_f \) and \( W_s \) denote weight matrices, \( b_f \) and \( b_s \) denote bias vectors, \( \sigma(\cdot) \) denotes the sigmoid function, \( \odot \) denotes the element-wise product.
element-wise multiplication, and \( g(\cdot) \) the softplus activate function.\(^8\)

The CGCNN employs not only the topological information of the crystal graph but also the spatial information of the distance features. However, the spatial information is actually unnecessary for predicting equilibrium properties. For molecular property predictions, the gated graph neural networks were used without the spatial information.\(^9,10\) As shown in Table 10 of Gilmer’s paper,\(^10\) for the atomization energy at 0 K and the electron energy gap, the errors of the no distance model are 0.72 kcal/mol and 97 meV comparable to ones of the distance model, 0.55 kcal/mol and 75 meV, respectively.

This study employs the crystal graph coordinator defined later, and demonstrates that the CGNN models can predict bulk properties with high precision.

This paper proposes the data mining system based on the CGNN as shown in Fig. 1. The crystal graph generator (CG-Gen) is a function of the atomic number sequence \( Z \), and sequentially produces the crystal graph of an equilibrium structure \( G = (Z,E) \) where \( E \) is the sequence of the neighbor lists \( \{N_i\}_i=1^N \) that defines the multiset of directed edges \( \{e_{ij} : j \in N_i\}_i=1^N \). The P-CGNN is a multitask model to predict the equilibrium properties \( P = (H_f, V_c, E_g, M_t, \ldots) \). Information about the unit cell volume \( V_c \) is included in the spatial information. The multitask model is composed of individual CGNN models dedicated to predicting only one property. The S-CGNN is a CGNN model to infer both the cell shape \( A_0 \) and the fractional coordinates \( X \) under the condition \( \det(A_0) = 1 \). We can obtain the equilibrium unit cell by scaling \( A_0 \) with the predicted unit cell volume as \( A = V_c^{1/3} A_0 \).

To train the CGNN and the CG-Gen, we need a set of crystal graphs made from the equilibrium structures. In crystallography, Voronoi diagrams are used to find the nearest neighbors of an atom.\(^11\) The CGCNN study employed this approach as mentioned in Supporting Information.\(^7\) This paper introduces the crystal graph coordinator that makes connections between atoms by clustering scaled interatomic distances, as described in the next section.

Figure 1: The data mining system for materials science.

**Algorithm 1** Crystal Graph Coordinator

**Parameters:** Radius cutoff factor \( \alpha = 1.2 \), Minimum distance between clusters \( \beta = 0.03 \)

**Require:** Lattice vectors \( A \), Sequence of atomic numbers \( \{z_i\}_i=1^N \), Sequence of fractional coordinates \( \{x_i\}_i=1^N \)

**Ensure:** Sequence of neighbor lists \( \{N_i\}_i=1^N \)

1: \( V_c \leftarrow \det(A) \) \( \triangleright \) Volume of the unit cell
2: \( r \leftarrow \{\rho(z_i)\}_i=1^N \) \( \triangleright \) Atomic radii
3: \( V_0 \leftarrow \sum_{i=1}^N \frac{1}{3} \pi r_i^3 \) \( \triangleright \) Total atomic volume
4: \( f_v \leftarrow \left(\frac{V_v}{V_c}\right)^{\frac{1}{3}} \) \( \triangleright \) Volume correction factor
5: for \( i \leftarrow 1, N \) do
6: \( k \leftarrow 1, J \leftarrow \phi, D \leftarrow \phi \)
7: for \( j \leftarrow 1, N \) do
8: \( c \leftarrow (r_i + r_j) f_v \alpha \) \( \triangleright \) Cutoff radii
9: for all \( n \in Z^3 \) do
10: \( d \leftarrow \|A(x_j - x_i + n)\| \)
11: if \( d < c \) then
12: \( J \leftarrow j, D_k \leftarrow d/c \)
13: \( k \leftarrow k + 1 \)
14: repeat
15: \( (C_1, C_2) \leftarrow \psi(D) \)
16: \( \Delta \leftarrow \mu(D[C_2]) - \mu(D[C_1]) \)
17: if \( \Delta < \beta \) then
18: break
19: else
20: \( J \leftarrow J[C_1], D \leftarrow D[C_1] \)
21: until \( |J| \leq 1 \)
22: return \( \{N_i\}_i=1^N \) \( \triangleright \) Neighbor list for \( i \)-th atom

This paper studies the system composed of the crystal graph coordinator and the CGNN as shown in Fig. 2. It remains for future work to integrate the CG-Gen and S-CGNN, as discussed later.

**2 The Crystal Graph Coordinator.**

This study employs the crystal graph coordinator described by Algorithm 1. The function \( \rho(\cdot) \) is a map from the atomic number to the atomic radius. The atomic radii are obtained from the PyMatGen library.\(^12\) The cutoff radius for \( i \)-th and \( j \)-th atoms \( c \) is mainly calculated by the sum of the two atomic radii, depends on the unit cell volume \( V_c \) through the multiplication of the
The convolution block is composed of the edge neural network (EdgeNet), the gated convolution, and the multilayer fully connected neural network (MFCNet), as shown in Fig. 3(b). The EdgeNet shown in Fig. 3(c) is given by

\[ s_{ij} = h_j W_e + f(h_i W_h) \odot f(h_j W_e), \]  

where the activation function \( f(\cdot) \) is the exponential linear unit (ELU),\(^{16} \) and \( W_e, W_h, W_e \in \mathbb{R}^{d_h \times d_h} \) denote weight matrices. The gated convolution is given by

\[ h_i^{t, \text{conv}} = \sum_{j \in N_i} \sigma(s_{ij} W_f) \odot (s_{ij} W_s), \]  

where \( W_f, W_s \in \mathbb{R}^{d_e \times d_h} \) denote weight matrices. When the EdgeNet disappears, \( d_e = d_h \) and \( s_{ij} = h_j, h_i^{t, \text{conv}} \) is fed to the MFCNet which consists of \( L_c \) fully connected layers that employ the ELU activation function after the batch normalization of the output.\(^{17} \) Each MFCNet layer uses a \( d_h \times d_h \) weight matrix. The next hidden state \( h_i^{(t+1)} \) is the sum of the shortcut hidden state \( h_i^{(t)} \) and the \( i \)-th MFCNet output.

The graph-level representation \( \Gamma^{(0)} \) is made from all the hidden states \( \{h_i^{(t)}\}_{t=1,i=1}^{T,N} \) except for the initial ones in accordance with the following procedure. At each step \( t \), the hidden states \( \{h_i^{(t)}\}_{t=1,i=1}^{T,N} \) are pooled with the gating mechanism as

\[ \gamma_t = \frac{1}{N} \sum_{i=1}^{N} \sigma(h_i^{(t)} W^{(t)}_\gamma + b^{(t)}_\gamma) \odot h_i^{(t)}, \]  

where \( W^{(t)}_\gamma \in \mathbb{R}^{d_h \times d_h} \) denotes a weight matrix, and \( b^{(t)}_\gamma \in \mathbb{R}^{d_h} \) is a bias vector. Then, the graph-level states \( \gamma_1, \ldots, \gamma_T \) are weightedly averaged as

\[ \Gamma^{(0)} = g \left( \sum_{t} \gamma_t W^{(t)}_\Gamma \right), \]  

where \( W^{(t)}_\Gamma \in \mathbb{R}^{d_h \times d_h} \) denotes a weight matrix. The batch normalization is applied before the softplus activation \( g(\cdot) \).

The initial graph state \( \Gamma^{(0)} \in \mathbb{R}^{d_h} \) is fed to the graph-level MFCNet which is almost the same as the convolution-block MFCNet. However, this MFCNet uses a \( d_h \times d_g \) weight matrix for the first fully connected layer to change the graph state size to \( d_g \), and consists of \( L_g \) layers that employ the softplus activation instead of the ELU activation. The \( l \)-th layer produces the graph-level output \( \Gamma^{(l)} \in \mathbb{R}^{d_g} \). The final layer’s output \( \Gamma^{(L_g)} \) is used as the input for the linear regression to predict the target value.

3 Crystal Graph Neural Networks

As shown in Fig. 3(a), the CGNN architecture is constructed starting with the embedding layer with respect to the atomic numbers \( \{z_i\}_{i=1}^{N} \) which gives the initial hidden states

\[ h_i^{(0)} = \mathcal{E}(z_i). \]  

The sequence of \( T \) convolution blocks produces a sequence of higher-level hidden states \( \{h_i^{(t)}\}_{t=1}^{T} \). The \( t \)-th hidden state \( h_i^{(t)} \) is fed to the \( (t+1) \)-th convolution block, which outputs the next-level hidden state \( h_i^{(t+1)} \). All the hidden states have the same size \( d_h \).

---

\(^{1} \) \( D[K] = \{D_{K_i}\}_{i=1}^{K} \) where \( K \) denotes an index list.

\(^{2} \) The X-Means algorithm also uses the 2-Means algorithm recursively.\(^^{14} \)
4 Experiments

4.1 Experimental Setup

This study uses a dataset composed of relaxed structures, formation energies, unit cell volumes, band gaps, and total magnetizations extracted from 561,888 distinct entries in the OQMD. The crystal graph for every extracted structure is created using the crystal graph coordinator with the parameters $\alpha = 1.2$ and $\beta = 0.03$ which are determined by trial and error. Some crystal graphs are shown in Appendix A.

There are 338,135 reduced chemical formulas in the dataset. Two distinct 10% random samples of the multi-component reduced formulas are reserved for the testing and validation sets, respectively. In other words, all unary formulas are excluded for the random samplings. The remaining formulas and all unary formulas are used for the training set. Consequently, the training, validation, and testing sets are composed of 449,915, 56,045, and 55,928 structures, respectively.

The hidden state size $d_h$ is restricted to $32n$ where $n \in [1 \ldots 7]$, and the edge and graph state sizes are proportional to the hidden state size as $d_e = \frac{3}{2}d_h$ and $d_g = 2d_h$. As shown in Fig. 3, $T = 4$ and $L_c = L_g = 2$, if not mentioned otherwise. The CGNN model without the EdgeNet or the convolution-block MFCNet for $d_h = 96$ serves as the benchmark model.

The author implemented the CGNN using the PyTorch library (v1.0). All networks are trained using the Adam optimization method with the batch size of 512. The optimizer uses a weight decay of $10^{-6}$ for regularization. The total number of epochs is 300 if not mentioned, and 600 for a few models. The learning rate is initially set to $10^{-3}$, and is reduced by multiplication of 0.1 at 5/6 of the total number of training epochs. The loss function is the mean of squared errors (MSE), while the validation metric is the mean of absolute errors (MAE). The final model employs the model weights that give the best validation metric. The testing evaluation uses the root of MSE (RMSE) and the MAE. Only the MAE is used to determine top models.

The insulating and magnetic probabilities of a material are calculated as

$$ p = \sigma(\zeta_{\text{train}}(y_{\text{test}})), $$

where $y_{\text{test}}$ is a predicted value of either band gap or total magnetization per atom, and $\zeta_{\text{train}}(\cdot)$ is the z-score function based on the training set. The ground-truth label is false if the target value is less than $10^{-2}$ eV for band gap and $\mu_H$/atom for total magnetization, and true otherwise. The area under the receiver operating characteristic curve (ROC-AUC) is used as a metric for classification problems with respect to insulating and magnetic materials.

The internal uncertainty of the OQMD for each property is measured in the form of RMSE and MAE metrics by comparison with the MP database (Appendix B). For each property, the database metrics are compared with the prediction metrics of the ensemble model comprised of a few top models. The ensemble prediction is calculated as the simple average of the member predictions.

<table>
<thead>
<tr>
<th>$L_c$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Benchmark)</td>
<td>0.0907</td>
<td>0.0436</td>
</tr>
<tr>
<td>1</td>
<td>0.0889</td>
<td>0.0417</td>
</tr>
<tr>
<td>2</td>
<td>0.0866</td>
<td>0.0414</td>
</tr>
<tr>
<td>3</td>
<td>0.0876</td>
<td>0.0414</td>
</tr>
</tbody>
</table>

Table 1: Formation energy prediction errors of the CGNN models without the EdgeNet for $d_h = 96$ are shown in eV. The convolution-block MFCNet vary in layer count ($L_c$) from 0 to 3.

<table>
<thead>
<tr>
<th>$d_h$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No EdgeNet</td>
<td>0.0880</td>
<td>0.0404</td>
</tr>
<tr>
<td>128</td>
<td>0.0869</td>
<td>0.0395</td>
</tr>
<tr>
<td>192</td>
<td>0.0855</td>
<td>0.0389</td>
</tr>
<tr>
<td>224</td>
<td>0.0869</td>
<td>0.0392</td>
</tr>
<tr>
<td>300 epochs</td>
<td>96</td>
<td>0.0855</td>
</tr>
<tr>
<td>128</td>
<td>0.0863</td>
<td>0.0366</td>
</tr>
<tr>
<td>192</td>
<td>0.0855</td>
<td>0.0365</td>
</tr>
<tr>
<td>600 epochs</td>
<td>128</td>
<td>0.0861</td>
</tr>
<tr>
<td>160</td>
<td>0.0857</td>
<td>0.0351</td>
</tr>
<tr>
<td>192</td>
<td>0.0848</td>
<td>0.0346</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.0794</td>
<td>0.0305</td>
</tr>
<tr>
<td>Database</td>
<td>0.1242</td>
<td>0.0848</td>
</tr>
</tbody>
</table>

Table 2: Formation energy prediction errors of the CGNN models are shown in eV. The ensemble model is comprised of the three models trained for 600 epochs.

The negative skewness of 1 eV/atom.

4.2 Formation Energy

We first examine results of a statistical analysis performed on the formation energy data in the testing set. The minimum and maximum formation energies are −4.42 and 4.46 eV/atom, respectively. The mean and the standard deviation are 0.08 and 0.86 eV/atom, respectively. The negative skewness of −1.26 is due to the second peak at about −1.9 eV/atom with a height that is 21 times lower than one of the first peak near the mean. The kurtosis of 3.70 suggests that the data is more heavy-tailed than a Laplace distribution,† but the large kurtosis is mainly due to the broad distribution in the range less than −1 eV/atom.

†This study employs the kurtosis definition by which the kurtosis is 0, 3, and 6 for normal, Laplace, and exponential distributions, respectively.

\[\text{https://github.com/Tony-Y/cgnn}\]
Table 3: Volume deviation prediction errors of the CGNN models are presented with a comparison between the OQMD and the MP database based on 27,057 matched entries. The ensemble model is comprised of the top four models.

<table>
<thead>
<tr>
<th>$d_h$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>96</td>
<td>0.0342</td>
</tr>
<tr>
<td>No EdgeNet</td>
<td>96</td>
<td>0.0339</td>
</tr>
<tr>
<td>300 epochs</td>
<td>96</td>
<td>0.0337</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.0315</td>
<td>0.0155</td>
</tr>
<tr>
<td>Database</td>
<td>0.0421</td>
<td>0.0270</td>
</tr>
</tbody>
</table>

Table 1 shows the performances of the CGNN models without the EdgeNet for $d_h = 96$. The model for $L_c = 2$ gives the best RMSE. The model for $L_c = 3$ gives the best MAE, but has only the slight lower MAE than the model for $L_c = 2$. Therefore, this study hereafter employs only the MFCNet with $L_c$ of 2.

The CGNN without the EdgeNet performs best for $d_h = 192$, as shown in the first section of Table 2. The CGNN trained for 300 epochs gives better MAE and almost the same RMSE for $d_h = 160$, as shown in the second section of Table 2. The CGNN trained for 600 epochs gives the best MAE and RMSE for $d_h = 192$, as shown in the third section of Table 2. The ensemble model is comprised of the three CGNN models trained for 600 epochs with $d_h$ of 128, 160, and 192, respectively, and gives much better MAE and RMSE than the database.

### 4.3 Unit Cell Volume

In this paper, the volume deviation is defined as

$$\delta_{vol} = 1 - \frac{V_a}{V_c},$$

where $V_a$ denotes the total atomic volume, and $V_c$ denotes the unit cell volume.

The minimum and maximum volume deviations in the testing set are $-2.00$ and $0.95$, respectively. The mean and the standard deviation are $-0.12$ and $0.34$, respectively. The volume deviation data has the skewness of $-0.01$ near zero and the positive kurtosis of 0.36 less than 1, and looks like a normal data.

The CGNN model without the EdgeNet performs slightly better than the benchmark model, using $d_h$ of 96 and 128, as shown in Table 3. A little improvement is seen for both the CGNN models with $d_h$ of 96 and 128 trained for 300 and 600 epochs, respectively. The ensemble model is comprised of these four models, and gives better MAE and RMSE than the database.

### 4.4 Band Gap

There are 52,993 metals and 2,935 insulators in the testing set. The metallic set is composed of almost zero band gaps which occupy 94.8% of the testing set, while the insulating set has the mean of 2.14 eV, the standard deviation of 1.75 eV, and the maximum of 9.87 eV. The band gap data is nonnegative and biased at zero. The insulating data has the skewness of 1.05 near unity and the kurtosis of 0.59 less than 1, and looks like a half-normal data in the range greater than 1 eV.

The benchmark model cannot successfully have been optimized due to less regularization. The use of the EdgeNet cannot improve the score for $d_h = 96$ as shown in Table 4. The CGNN model without the EdgeNet performs best for $d_h = 128$. For the metal sample, it gives slightly worse RMSE and MAE than the best ones given by the models with $d_h$ of 96 and 128, respectively. For the insulator sample, it gives the best MAE and RMSE. The ensemble model is comprised of the top three CGNN models without the EdgeNet and with $d_h$ of 96, 128, and 160, respectively, and gives better MAE and RMSE than the database. For the metal sample, it gives much better RMSE and MAE than the database. For the insulator sample, it gives worse RMSE and MAE than the database.

Table 4: Band gap prediction errors of the CGNN models are shown in eV. The ensemble model is comprised of the top three models.

<table>
<thead>
<tr>
<th>$d_h$</th>
<th>Total</th>
<th>Metal</th>
<th>Insulator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>No EdgeNet</td>
<td>32</td>
<td>0.2806</td>
<td>0.0580</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>0.2859</td>
<td>0.0546</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>0.2743</td>
<td>0.0516</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.2689</td>
<td>0.0511</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.2801</td>
<td>0.0516</td>
</tr>
<tr>
<td>Ensemble</td>
<td>96</td>
<td>0.3086</td>
<td>0.0565</td>
</tr>
<tr>
<td>Database</td>
<td>0.2556</td>
<td>0.0461</td>
<td>0.1476</td>
</tr>
<tr>
<td></td>
<td>0.5288</td>
<td>0.1806</td>
<td>0.3568</td>
</tr>
</tbody>
</table>
4.5 Total Magnetization

There are 30,299 nonmagnets and 25,629 magnets in the testing set. The nonmagnetic set consists of almost zero total magnetizations which occupy 54.2% of the testing set, while the magnetic set has the mean of 0.68 $\mu_B$/atom, the standard deviation of 0.58 $\mu_B$/atom, and the maximum of 4.68 $\mu_B$/atom. The total magnetization data is nonnegative and biased at zero like the band gap data, but the magnetic data has the positive skewness of 1.66 greater than 1 and the high kurtosis of 4.72, which suggests that it is very heavy-tailed, and looks like an exponential data.

As shown in Table 6, for $d_h = 160$, the CGNN model without the EdgeNet gives the best RMSE for all total, nonmagnet, and magnet samples, while the complete CGNN model gives the best MAE for all the samples. The ensemble model is comprised of the top three CGNN models with $d_h$ of 96, 128, and 160, respectively, determined only by the total MAE metric, and gives better metrics than the database except for the nonmagnet MAE.

Fig. 4 shows the ROC curve of the ensemble prediction for magnet. The ensemble model gives much better ROC-AUC than the database as shown in Table 5, and classifies metal and insulator examples in a high precision.

5 Discussions

5.1 High-throughput Screening

The OQMD has been employed as a source of data searched for candidate materials on the basis of high-throughput DFT. We can use the convex hull to extract the metastable candidates that could appear at the room temperature $T_{room}$. The formation energy of a

---

Table 6: Total magnetization prediction errors of the CGNN models are shown in $\mu_B$/atom. The ensemble model is comprised of the top three models.

<table>
<thead>
<tr>
<th>$d_h$</th>
<th>RMSE</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Nonmagnet</td>
<td>Magnet</td>
<td></td>
<td>Nonmagnet</td>
<td>Magnet</td>
</tr>
<tr>
<td>Benchmark</td>
<td>96</td>
<td>0.2025</td>
<td>0.0856</td>
<td>0.1243</td>
<td>0.0358</td>
<td>0.2669</td>
</tr>
<tr>
<td>No EdgeNet</td>
<td>96</td>
<td>0.2006</td>
<td>0.0827</td>
<td>0.1181</td>
<td>0.0333</td>
<td>0.2671</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.2010</td>
<td>0.0807</td>
<td>0.1242</td>
<td>0.0328</td>
<td>0.2644</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.1975</td>
<td>0.0793</td>
<td><strong>0.1196</strong></td>
<td>0.0322</td>
<td><strong>0.2612</strong></td>
</tr>
<tr>
<td>Ensemble</td>
<td>96</td>
<td>0.2023</td>
<td>0.0773</td>
<td>0.1226</td>
<td>0.0294</td>
<td>0.2675</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.2003</td>
<td>0.0764</td>
<td>0.1209</td>
<td>0.0292</td>
<td>0.2652</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>0.1997</td>
<td><strong>0.0753</strong></td>
<td>0.1228</td>
<td><strong>0.0284</strong></td>
<td>0.2631</td>
</tr>
<tr>
<td>Database</td>
<td>0.1855</td>
<td>0.0691</td>
<td>0.1147</td>
<td>0.0266</td>
<td><strong>0.2440</strong></td>
<td><strong>0.1194</strong></td>
</tr>
<tr>
<td></td>
<td>0.4003</td>
<td>0.0938</td>
<td>0.1399</td>
<td><strong>0.0211</strong></td>
<td>0.7824</td>
<td>0.3274</td>
</tr>
</tbody>
</table>

---

1 In this paper, antiferromagnetic materials are classified as nonmagnets because their total magnetizations are zero.
material is greater than or equal to the convex hull energy at its composition. If the energy above hull (EaH) of a material is lower than 50 meV ≈ 2k_B T_{room}, it is highly possible that the material is metastable at T_{room}.

The hull energy at every composition in the testing set was calculated on the basis of the training set using the phase diagram module in the PyMatGen package. The metastable set is composed of 6,306 materials that are in the testing set and have a target EaH less than 50 meV/atom. The sets of materials with predicted EaH less than 96 and 215 meV/atom include 95% and 99% of the metastable set, and are 1.26 and 2.27 times larger than the metastable set, respectively. The threshold of 100 meV for predicted EaH would work for extracting candidates in a high-throughput manner.

However, the CGNN ensemble models cannot be used in practical applications yet because we have no CG-Gen model that is necessary for generating crystal graphs of equilibrium structures. The next subsection discusses the data mining system including this point.

5.2 Materials Data Mining

Data mining systems for materials science would be desired to discover novel materials. The proposed system is constructed from the CG-Gen, P-CGNN, and S-CGNN along with the crystal graph coordinator (Fig. 1 and 2).

The crystal graph coordinator extracts topological information of a given structure in the form of crystal graph. The created graph expresses only the strong connections between every atom in the unit cell and its near neighbor atoms, and can sometimes be a disjoint union of connected subgraphs (see Appendix A, Fig. 5(d)). The S-CGNN must infer the unit cell shape and the fractional coordinates only from the graph. The graph separation is an obstacle to the placement of the separated groups of atoms. To resolve this problem, we can extend the crystal graph by adding some extra edges between the unconnected neighbors, that is, by appending an extra graph expressing the secondary structure between the separated groups.\(^1\)

The S-CGNN has not only the graph-level output expressing the unit cell shape, but also the node-level outputs expressing the fractional coordinates. Although there are no node-level outputs in the CGNN architecture proposed in this paper, we can extend the architecture in diverse ways. The addition of the gated pooling with respect to every node to the CGNN architecture would be the most straight extension, but we have to investigate whether it works as expected.

Suppose that the CG-Gen generates the final graph \(G^{(M)} = (Z, E)\) through \(M\) transformations beginning with the initial graph \(G^{(0)} = (Z, \phi)\), where \(E\) denotes the final multiset of edges and \(\phi\) denotes an empty set. We can consider at least three graph transformations which adds, deletes, and replaces an edge, respectively.\(^2\) One of them is applied to the graph at each step. If we can obtain the graph sequence \(\{G^{(m)}\}_{m=0}^M\) approaching an equilibrium crystal graph, recurrent graph neural networks can be used to learn this graph sequence.\(^3\) For example, if only three edge transformations of addition, deletion, and inaction are considered, the transforming probability of the edge \(e_{ij}\) can be calculated as the softmax transformation of the 3-dimensional output of a neural network \(T(h_i, h_j)\) where \(h_i\) and \(h_j\) denote the \(i\)-th and \(j\)-th hidden states in a recurrent graph neural network, respectively.

Through the domain knowledge of materials science, we can see such sequences in nature. When a few isolated atoms come together, they may be connected with chemical bonds and transform into a molecule. A larger molecule is synthesized from a couple of molecules. When there is a crystal nucleus in a liquid cooled below its freezing point, the nucleus grows attracting atoms in the liquid. In a vacuum chamber, vapor molecules are deposited on a substrate and solid materials are made from them. This method to make high-quality materials is well-known as chemical vapor deposition.

Annealing makes the material progress towards its other equilibrium state because of the thermally activated diffusion of atoms. We can therefore consider some graph sequence starting with an equilibrium graph and ending with another equilibrium graph. Nevertheless, the intermediate graphs in the sequence need not be restricted to crystal graphs for existing nonequilibrium structures. Since we cannot have the complete set of the intermediate graphs, we have to train the CG-Gen in a semi-supervised fashion. Because the CG-Gen must output only equilibrium graphs sequentially, assigning a label to every crystal graph in the sequence according to whether or not it is an equilibrium graph, we have to resolve the equilibrium classification problem.

The materials data mining would be realized when the CG-Gen almost completely learns synthetic chemistry on crystal graphs from an appropriate materials database. Since a crystal graph sequence is generated by the corresponding graph transformation sequence, an entire set of graph transformation sequences could be considered as the materials genome that stores complete information of materials synthesis, like the human genome is the complete set of nucleic acid sequences for humans. It is not difficult to predict equilibrium properties from a proper crystal graph produced by the materials genome, as demonstrated in this paper.

6 Conclusions

For every material in the OQMD testing set, the ensemble CGNN models have successfully predicted the formation energy, the unit cell volume, the band gap, and the

\(^1\) In molecular biology, the secondary structure is the pattern of hydrogen bonds in a biopolymer.

\(^2\) In materials science, the secondary structure is the pattern of chemical bonds in a crystal.

\(^3\) The softmax transformation is a function that takes as input a vector of real numbers and returns a vector of values between 0 and 1 that sum to 1. It is often used to convert a vector of real numbers to a probability distribution.
total magnetization, and classified it under metal and insulator, and under nonmagnet and magnet. It is important that the predictions are given without the use of spatial information. The materials data mining would become sophisticated when the crystal graph generator that can learn the synthetic chemistry is constructed using the recurrent neural networks based on the CGNN architectures.

Acknowledgment

I would like to thank Christopher Sutton for brief discussions on CGNN models that I have created for a few months after the summer school where I presented the 1st place solution in the NOMAD2018 Kaggle competition.26

References


A Crystal Graph Examples

Fig. 5 shows crystal graph examples.

(a) (b) (c) (d)

Figure 5: The crystal graphs for (a) diamond cubic Si, (b) α-SiO\textsubscript{2}, (c) Mg\textsubscript{3}Al\textsubscript{2}(SiO\textsubscript{4})\textsubscript{3} garnet known as pyrope, and (d) Li\textsubscript{3}Nd\textsubscript{3}(WO\textsubscript{6})\textsubscript{2} garnet.
Comparison of the OQMD with the MP Database

The OQMD and MP use the same software to calculate electronic structures, ionic configurations, and equilibrium properties. There are substantial similarities between the OQMD and the MP database. Therefore, a calculation in the OQMD is comparable to the corresponding calculation in the MP database.

Matched structures of the OQMD and the MP database for each chemical formula were found using the StructureMatcher class included in the PyMatGen package. There are 27,074 matched materials, each of which is the most stable for its reduced chemical formula.

The relative volume error between the OQMD and the MP entry is defined as

$$\delta V_{DB} = 2\frac{V_{OQMD} - V_{MP}}{V_{OQMD} + V_{MP}}, \quad (11)$$

where $V_{OQMD}$ and $V_{MP}$ denote the unit cell volume of the OQMD and the MP entry, respectively. The histogram of relative volume errors is shown in Fig. 6. This study uses 27,057 inliers within between the lower and the upper bound depicted as the red vertical line. There are 5 lower and 12 upper outliers, each of which is shown as a solid half circle on the bottom axis that is annotated with its reduced chemical formula.

The property error between the OQMD and the MP entry is calculated by

$$\epsilon = X_{OQMD} - X_{MP}, \quad (12)$$

where $X_{OQMD}$ and $X_{MP}$ denote the OQMD and the MP property, respectively. The histograms of errors for formation energy, volume deviation, band gap, and total magnetization are shown in Fig. 7, 8, 9, and 10, respectively. The RMSE and MAE for each property are shown in Table 7. A material is determined to be metallic if its band gap in the MP database is less than 0.01 eV, and insulating otherwise. A material is also determined to be nonmagnetic if its total magnetization in the MP database is less than 0.01 $\mu_B$/atom, and magnetic otherwise. The insulating and magnetic probabilities are calculated using Eq. (9), but the z-score function is based on the MP database. The ROC curves for insulator and magnet classifications of the OQMD are shown in Fig. 11.

The database errors are ascribed to neither the DFT nor the quantum chemical simulation package, but come from the difference in strategy between the OQMD and the MP database. The analysis of the database uncertainty would encourage improvement in the strategies of theoretical databases.
Figure 9: The histograms of OQMD and MP band gaps (left), and OQMD-MP band gap errors (right).

Figure 10: The histograms of OQMD and MP total magnetizations (left), and OQMD-MP total magnetization errors (right).

Table 7: Errors of formation energies, volume deviations, band gaps, and total magnetizations between the OQMD and the MP database.

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formation energy (eV/atom)</td>
<td>0.1242</td>
<td>0.0848</td>
</tr>
<tr>
<td>Volume deviation</td>
<td>0.0421</td>
<td>0.0270</td>
</tr>
<tr>
<td>Band gap (eV)</td>
<td>0.5288</td>
<td>0.1806</td>
</tr>
<tr>
<td>For 14723 metals</td>
<td>0.3568</td>
<td>0.0461</td>
</tr>
<tr>
<td>For 12334 insulators</td>
<td>0.6794</td>
<td>0.3412</td>
</tr>
<tr>
<td>Total magnetization ($\mu_B$/atom)</td>
<td>0.4003</td>
<td>0.0938</td>
</tr>
<tr>
<td>For 20632 nonmagnets</td>
<td>0.1399</td>
<td>0.0211</td>
</tr>
<tr>
<td>For 6425 magnets</td>
<td>0.7824</td>
<td>0.3274</td>
</tr>
</tbody>
</table>

Figure 11: The receiver operating characteristic curves for insulator and magnet classifications when comparing the OQMD with the MP database.