Crystal 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 testing 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 metal-insulator and nonmagnet-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 is defined by lattice vectors $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) \in \mathbb{R}^{3\times3}$, fractional coordinates $X = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{F}^{N\times3}$, and atomic numbers $Z = \{z_i\}_{i=1}^{N} \in \mathbb{I}^{N}$, where $\mathbb{F} = \{x \in \mathbb{R} : 0 \leq x < 1\}$, and $\mathbb{I} = [1\ ..\ 118]$. The Cartesian position of i-th atom is calculated as $A\mathbf{x}_i$ in the unit cell defined by A. The unit cell volume $V_c = \det(A)$ must be greater than 0. A macroscopic crystal is made of the unit cells stacked in three dimensions.

The total energy of a crystalline material depends on its structure $S = (A, X, Z)$, 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;

$$\mathfrak{S} = \arg\min_{S} E_{\text{tot}}(S) \text{ subject to } Z(S) \text{ fixed.} \tag{1}$$

There are multiple minima, and $E_{\text{tot}}(\mathfrak{S})$ 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, $\odot$

*yamamoto.takenory@gmail.com

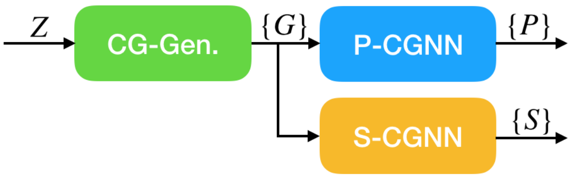


Figure 1: The data mining system for materials science.

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. The former errors are approximately 1.3 times the latter ones. These results suggest that the distance features are inessential to predicting the equilibrium properties. My study does not use any spatial information in crystal graph neural networks (CGNNs) 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 $\{\mathcal{N}_i\}_{i=1}^N$ that defines the multiset of directed edges $\{e_{ij} : j \in \mathcal{N}_i\}_{i=1}^N$. The P-CGNN is a multitask model to predict the equilibrium properties $P = (H_f, V_c, E_g, M_t, \dots)$. 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.

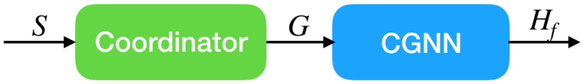


Figure 2: The crystal graph coordinator and the crystal graph neural network.

Algorithm 1 Crystal Graph Coordinator

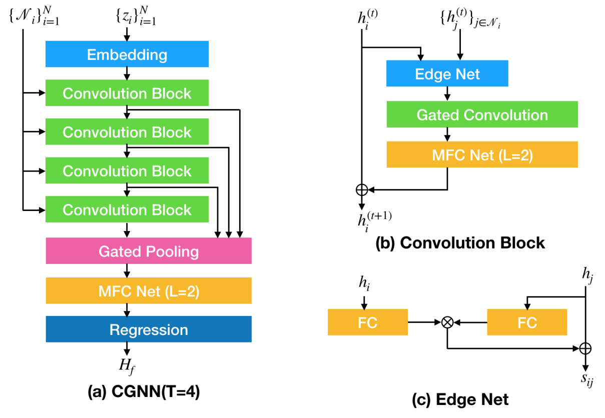
Parameters: Radius cutoff factor $\alpha = 1.2$, Minimum distance between clusters $\beta = 0.03$

```
Require: Lattice vectors $A$, Sequence of atomic numbers $\{z_i\}_{i=1}^N$, Sequence of fractional coordinates $\{\mathbf{x}_i\}_{i=1}^N$
Ensure: Sequence of neighbor lists $\{\mathcal{N}_i\}_{i=1}^N$
$V_c \leftarrow \det(A)$                                   ▷ Volume of the unit cell
$r \leftarrow \{\rho(z_i)\}_{i=1}^N$                      ▷ Atomic radii
$V_a \leftarrow \sum_{i=1}^N \frac{4\pi}{3} r_i^3$          ▷ Total atomic volume
$f_v \leftarrow \left(\frac{V_c}{V_a}\right)^{\frac{1}{3}}$   ▷ Volume correction factor
for $i \leftarrow 1, N$ do
    $k \leftarrow 1, J \leftarrow \phi, D \leftarrow \phi$
    for $j \leftarrow 1, N$ do
        $c \leftarrow (r_i + r_j) f_v \alpha$              ▷ Cutoff radii
        for all $\mathbf{n} \in \mathbb{Z}^3$ do
            $d \leftarrow \|A(\mathbf{x}_j - \mathbf{x}_i + \mathbf{n})\|$
            if $d < c$ then
                $J_k \leftarrow j$, $D_k \leftarrow d/c$
                $k \leftarrow k + 1$
    repeat
        $(C_1, C_2) \leftarrow \psi(D)$
        $\Delta \leftarrow \mu(D[C_2]) - \mu(D[C_1])$
        if $\Delta < \beta$ then
            break
        else
            $J \leftarrow J[C_1]$, $D \leftarrow D[C_1]$
    until $|J| \leq 1$
    $\mathcal{N}_i \leftarrow J$                          ▷ Neighbor list for $i$-th atom
return $\{\mathcal{N}_i\}_{i=1}^N$
```

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



(a) CGNN(T=4)

(b) Convolution Block

(c) Edge Net

Figure 3: The crystal graph neural network architecture.

volume correction factor $(V_c/V_a)^{1/3}$ where V_a is the total atomic volume, and is extended by multiplication of the radius cutoff factor $\alpha > 1$. The index list of atoms neighboring to the i-th atom J is generated by collecting atoms in the sphere with the radius of c centered at the i-th atom. Meanwhile, the distance between the i-th and j-th atom d is divided by c, and appended to the list D. The function $\psi(\cdot)$ gives two lists of indices C_1 and C_2 to split the list of scaled distances D into two clusters $D[C_1]$ and $D[C_2]$ using the K-Means algorithm $(K = 2)$,[13†] and ensures $\mu(D[C_1]) \leq \mu(D[C_2])$ where the function $\mu(\cdot)$ produces the mean of values in a given list. The clustering is repeated until the cluster distance Δ is less than the miminum distance $\beta < 1$.[‡] Moreover, if the number of neighbors $|J|$ equals to 1 after updating J, the next step is not executed. The above procedure is applied to every atom, and thus we obtain the neighbor lists for all atoms. The crystal graph is invariant under any uniform deformation of the unit cell because of the use of both the volume correction factor and the scaled distances.

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 $\{z_i\}_{i=1}^N$,[15] which gives the initial hidden states

$$h_i^{(0)} = \mathcal{E}(z_i). \tag{4}$$

The sequence of T convolution blocks produces a sequence of higher-level hidden states $\{h_i^{(t)}\}_{t=1}^T$. The level-t 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.

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

$$s_{ij} = h_j W_c + f(h_i W_h) \odot f(h_j W_e), \tag{5}$$

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

$$h_i^{\mathrm{conv}} = \sum_{j\in\mathcal{N}_i} \sigma(s_{ij} W_f) \odot (s_{ij} W_s), \tag{6}$$

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^{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 $\{h_i^{(t)}\}_{i=1}^N$ are pooled with the gating mechanism as

$$\gamma_t = \frac{1}{N}\sum_{i=1}^N \sigma(h_i^{(t)} W_\gamma^{(t)} + b_\gamma^{(t)}) \odot h_i^{(t)}, \tag{7}$$

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

$$\Gamma^{(0)} = g(\sum_t \gamma_t W_\Gamma^{(t)}), \tag{8}$$

where $W_\Gamma^{(t)} \in \mathbb{R}^{d_h \times d_h}$ 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.

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

[‡] 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 multicomponent 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 $n \in [1 \mathrel{..} 7]$, 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).[18‡] All networks are trained using the Adam optimization method with the batch size of 512.[19] 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

$$p = \sigma(\zeta_{\text{train}}(y_{\text{test}})), \tag{9}$$

where y_{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 μ_{B}/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.

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.

L_c	RMSE	MAE
0 (Benchmark)	0.0907	0.0436
1	0.0889	0.0417
2	**0.0866**	0.0414
3	0.0876	**0.0414**

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.

	d_h	RMSE	MAE
No EdgeNet	128	0.0880	0.0404
	160	0.0869	0.0395
	192	**0.0855**	**0.0389**
	224	0.0869	0.0392
300 epochs	96	**0.0855**	0.0370
	128	0.0863	0.0366
	160	0.0855	**0.0365**
600 epochs	128	0.0861	0.0357
	160	0.0857	0.0351
	192	**0.0848**	**0.0346**
Ensemble		**0.0794**	**0.0305**
Database		0.1242	0.0848

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.

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.

[‡] https://github.com/Tony-Y/cgnn

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

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.

	d_h	RMSE	MAE
Benchmark	96	0.0342	0.0181
No EdgeNet	96	0.0339	0.0178
	128	0.0337	0.0178
300 epochs	96	0.0337	0.0171
600 epochs	128	**0.0336**	**0.0170**
Ensemble		**0.0315**	**0.0155**
Database		0.0421	0.0270

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

$$\delta_{\text{vol}} = 1 - \frac{V_a}{V_c}, \tag{10}$$

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

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

		Total		Metal		Insulator	
	d_h	RMSE	MAE	RMSE	MAE	RMSE	MAE
No EdgeNet	32	0.2806	0.0580	0.1704	0.0250	0.9877	0.6550
	64	0.2859	0.0546	0.1685	0.0228	1.0220	0.6288
	96	0.2743	0.0516	**0.1585**	0.0211	0.9898	0.6033
	128	**0.2689**	**0.0511**	0.1597	0.0210	**0.9578**	**0.5957**
	160	0.2801	0.0516	0.1592	**0.0207**	1.0187	0.6083
	96	0.3086	0.0565	0.1824	0.0232	1.1017	0.6563
Ensemble		**0.2556**	**0.0461**	**0.1476**	**0.0178**	0.9229	0.5569
Database		0.5288	0.1806	0.3568	0.0461	**0.6794**	**0.3412**

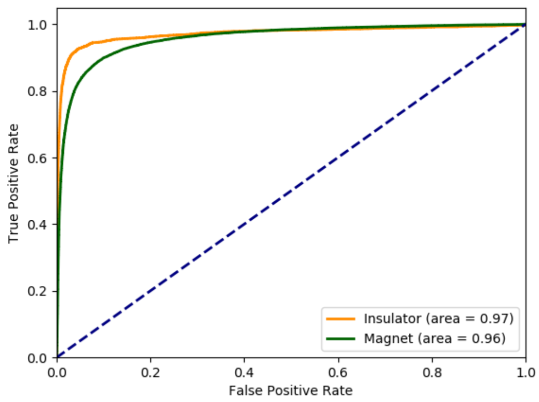


Figure 4: The receiver operating characteristic curves for the insulator and magnet classifications.

Table 5: The ROC-AUC for insulator and magnet classifications. The means and standard deviations used in the z-score functions are also listed.

Problem	Model	Mean	Std	AUC
Insulator	Ensemble	0.1166	0.6385	**0.9713**
	Database	1.0714	1.6023	0.9564
Magnet	Ensemble	0.3088	0.5200	**0.9569**
	Database	0.1519	0.4424	0.8688

database. These worse scores are probably due to the fewness of insulator examples.

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

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 μ_B/atom, the standard deviation of 0.58 μ_B/atom, and the maximum of 4.68 μ_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.[20, 21] We can use the convex hull to extract the metastable candidates that could appear at the room temperature T_room.[22] The formation energy of a

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

		Total		Nonmagnet		Magnet	
	d_h	RMSE	MAE	RMSE	MAE	RMSE	MAE
Benchmark	96	0.2025	0.0856	0.1243	0.0358	0.2669	0.1445
No EdgeNet	96	0.2006	0.0827	0.1181	0.0333	0.2671	0.1411
	128	0.2010	0.0807	0.1242	0.0328	0.2644	0.1374
	160	**0.1975**	0.0793	**0.1196**	0.0322	**0.2612**	0.1348
	96	0.2023	0.0773	0.1226	0.0294	0.2675	0.1340
	128	0.2003	0.0764	0.1209	0.0292	0.2652	0.1323
	160	0.1997	**0.0753**	0.1228	**0.0284**	0.2631	**0.1307**
Ensemble		**0.1855**	**0.0691**	**0.1147**	0.0266	**0.2440**	**0.1194**
Database		0.4003	0.0938	0.1399	**0.0211**	0.7824	0.3274

[†] 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 $\approx 2k_{\mathrm{B}}T_{\mathrm{room}}$, 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.[†]

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 ϕ denotes an empty set. We can consider at least three graph transformations which adds, deletes, and replaces an edge, respectively.[23] 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.[24,25] 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 $\mathcal{T}(h_i, h_j)$ 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

[†] In molecular biology, the secondary structure is the pattern of hydrogen bonds in a biopolymer.

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

[1] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15 – 50, 1996.

[2] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, Oct 1996.

[3] Anubhav Jain, Geoffroy Hautier, Charles J. Moore, Shyue Ping Ong, Christopher C. Fischer, Tim Mueller, Kristin A. Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, 50(8):2295 – 2310, 2011.

[4] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[5] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(11):1501–1509, Nov 2013.

[6] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *Npj Computational Materials*, 1:15010, Dec 2015.

[7] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018.

[8] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress.

[9] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. *CoRR*, abs/1511.05493, 2015.

[10] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *CoRR*, abs/1704.01212, 2017.

[11] V. A. Blatov, A. P. Shevchenko, and V. N. Serenzhkin. Crystal space analysis by means of Voronoi–Dirichlet polyhedra. *Acta Crystallographica Section A*, 51(6):909–916, Nov 1995.

[12] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314 – 319, 2013.

[13] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[14] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. 01 2002.

[15] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, Jan 2017.

[16] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[18] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[20] Antoine A. Emery and Chris Wolverton. High-throughput dft calculations of formation energy, stability and oxygen vacancy formation energy of abo3 perovskites. *Scientific Data*, 4:170153, Oct 2017.

[21] Muratahan Aykol, Soo Kim, Vinay I. Hegde, Scott Kirklin, and Chris Wolverton. Computational evaluation of new lithium-3 garnets for lithium-ion battery applications as anodes, cathodes, and solid-state electrolytes. *Phys. Rev. Materials*, 3:025402, Feb 2019.

[22] Wenhao Sun, Stephen T. Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D. Richards, Anthony C. Gamst, Kristin A. Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2(11), 2016.

[23] Hans-Jörg Kreowski, Renate Klempien-Hinrichs, and Sabine Kuske. Some essentials of graph transformation. In *Recent Advances in Formal Languages and Applications (Studies in Computational Intelligence)*, pages 229–254. Springer, 2006.

[24] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting. *CoRR*, abs/1709.04875, 2017.

[25] Jia Wang, Vincent W. Zheng, Zemin Liu, and Kevin Chen-Chuan Chang. Topological recurrent neural network for diffusion prediction. *CoRR*, abs/1711.10162, 2017.

[26] Takenori Yamamoto. The n-gram approach and its application to materials science, Sep 2018. NOMAD SUMMER 2018 (http://meetings.nomad-coe.eu/nomad-summer-2018).

A Crystal Graph Examples

Fig. 5 shows crystal graph examples.

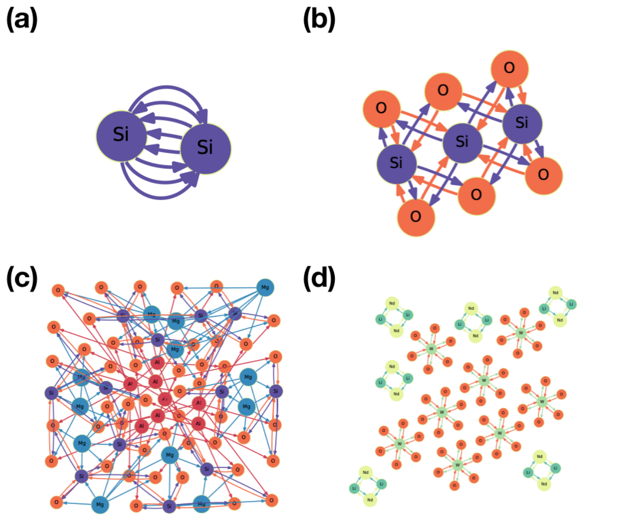


Figure 5: The crystal graphs for (a) diamond cubic Si, (b) α-SiO_2, (c) $Mg_3Al_2(SiO_4)_3$ garnet known as pyrope, and (d) $Li_3Nd_3(WO_6)_2$ garnet.

B 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

$$\delta V_{\mathrm{DB}} = 2\frac{V_{\mathrm{OQMD}} - V_{\mathrm{MP}}}{V_{\mathrm{OQMD}} + V_{\mathrm{MP}}}, \tag{11}$$

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

$$\epsilon = X_{\mathrm{OQMD}} - X_{\mathrm{MP}}, \tag{12}$$

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 metalic 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 μ_{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.

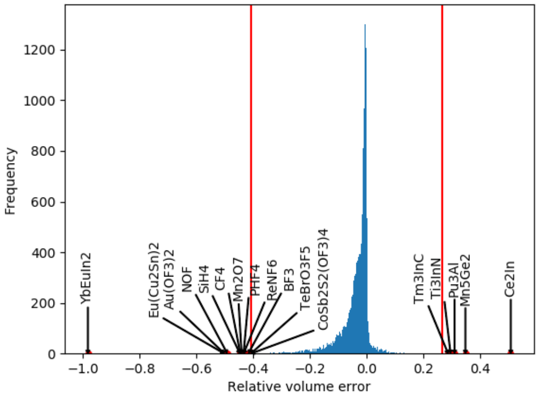


Figure 6: The histogram of OQMD-MP relative volume errors.

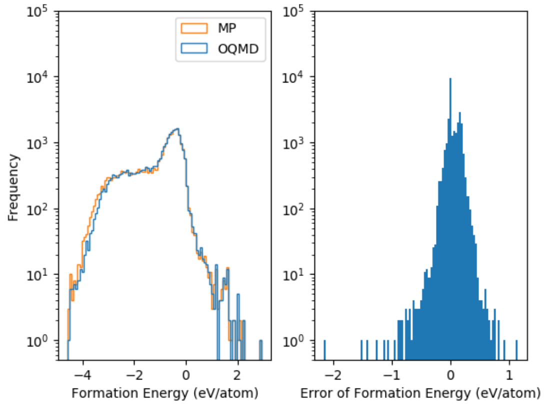


Figure 7: The histograms of OQMD and MP formation energies (left), and OQMD-MP formation energy errors (right).

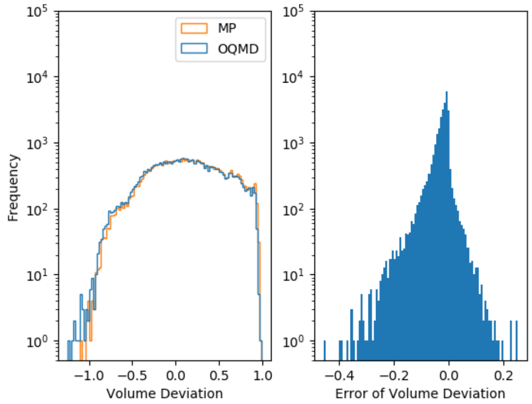


Figure 8: The histograms of OQMD and MP volume deviations (left), and OQMD-MP volume deviation errors (right).

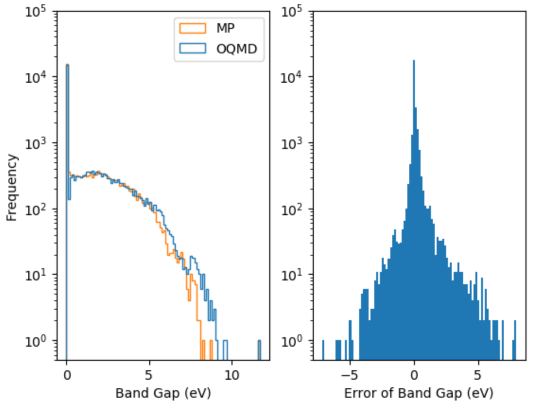


Figure 9: The histograms of OQMD and MP band gaps (left), and OQMD-MP band gap errors (right).

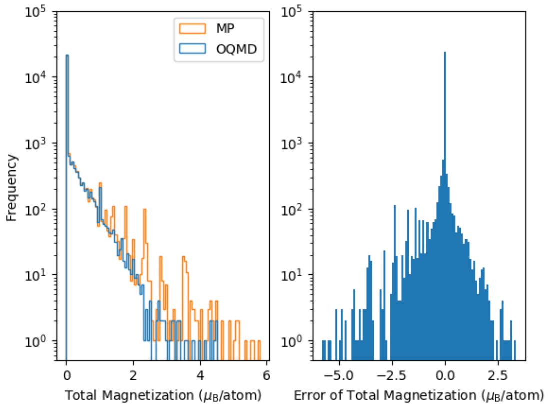


Figure 10: The histograms of OQMD and MP total magnetizations (left), and OQMD-MP total magnetization errors (right).

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

	RMSE	MAE
Formation energy (eV/atom)	0.1242	0.0848
Volume deviation	0.0421	0.0270
Band gap (eV)	0.5288	0.1806
For 14723 metals	0.3568	0.0461
For 12334 insulators	0.6794	0.3412
Total magnetization (μ_B/atom)	0.4003	0.0938
For 20632 nonmagnets	0.1399	0.0211
For 6425 magnets	0.7824	0.3274

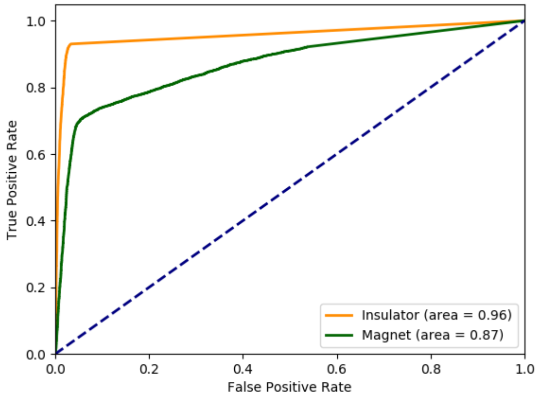


Figure 11: The receiver operating characteristic curves for insulator and magnet classifications when comparing the OQMD with the MP database.