OQM9HK: A Large-Scale Graph Dataset for Machine Learning in Materials Science

Takenori Yamamoto *

RIMCS LLC, Yokohama, Japan

September 30, 2022

Abstract

We introduce a large-scale dataset of quantum-mechanically calculated properties of crystalline materials for graph representation learning that contains approximately 900k entries (OQM9HK). This dataset is constructed on the basis of the Open Quantum Materials Database (OQMD) v1.5 containing more than one million entries, and the successor to the OQMD v1.2 dataset containing approximately 600k entries (OQM6HK). We develop the graph creation algorithm to produce a binary edge-labeled (BEL) graph representing a crystalline material. The BEL graph has higher representability of crystal structure than the edge-unlabeled ones. In materials property prediction tasks, crystal graph neural networks trained on the BEL graph dataset perform better than ones on the other graph datasets. The OQM9HK graph dataset is available at the Zenodo repository, https://doi.org/10.5281/zenodo.7124330

1 Introduction

Graphs are widely used to represent relationships between individuals, for examples, social and citation networks, etc. The citation graphs, CORA, CITESEER, and PUBMED, are used to evaluate node classification performances of machine learning models.[1] Recently, collections of midium-scale and large-scale graph datasets were developed for reliable benchmarking.[2–4]

We present a large-scale graph dataset of materials science based on the Open Quantum Materials Database (OQMD), which is a database of DFT calculated thermodynamic and structural properties of more than one million materials.[5] Before the OQMD being established, the Materials Project (MP) had developed a similar database since 2011.[6] The MP database contained approximately 30k materials as of Dec 2012. The current database (V2021.05.13) contains approximately 145k ones. The uncertainties of the OQMD and MP database for formation energy were estimated to be ∼100 meV/atom in the mean absolute error (MAE) by comparing calculated values with experimental ones.[5] Those databases have been used for high throughput screening of candidate materials.[7] However, materials scientists are frantically seeking for novel materials uncontained in materials databases.

Machine learning methods have substantially expedited researches in other scientific fields.[8–13] Especially, artificial intelligence caused a revolution in computational biology. The state-of-the-art protein structure predictor, Alphafold, was developed using a training dataset consisting of approximately 140k experimentally determined protein structures,[9] and then has built a protein structure database containing over 200 million entries.[10] In the future, a similar successful story could take place in materials science. Therefore, machine learning methods would be able to be applied to develop data mining systems that recommend new materials having some desirable properties. The MP recently introduced the machine learning benchmarking suite for materials science, named Matbench.[14] Matbench consists of 13 small-medium size ($\lesssim$ 100k) datasets for diverse tasks. The input of each dataset is chemical composition or crystal structure. Unfortunately, there is no structure prediction task in Matbench.

We proposed the crystal graph neural networks (CGNN) to predict the formation energy, unit cell volume, band gap, and total magnetization of a crystalline material, and showed that the CGNN models perform well on the graph dataset consisting of approximately 600k entries constructed on the basis of the OQMD v1.2 released in June 2018.[15†] Hereafter, we call the OQMD v1.2 dataset OQM6HK. We can elicit topological and spatial information from a crystal structure. The topological information is present as a graph expressing interatomic connections, that is, a crystal graph, while the spatial information may be given by unit cell volume or interatomic distances. The OQM6HK dataset contains the volume prediction task because the crystal graph is independent on the unit cell volume. The volume prediction task employs as the target the volume deviation, whose definition appears later (Eq. 1 in §2), instead of the unit cell volume. We estimated the database uncertainty by comparing every structurally matched pair of OQMD and MP database's entries. The MAEs of the ensemble models for the prediction of formation energy, volume deviation, band gap of insulator, and total magnetization of magnet

*yamamoto.takenory@gmail.com

†The OQMD v1.2 or OQM6HK graph dataset is available at https://doi.org/10.5281/zenodo.7118055

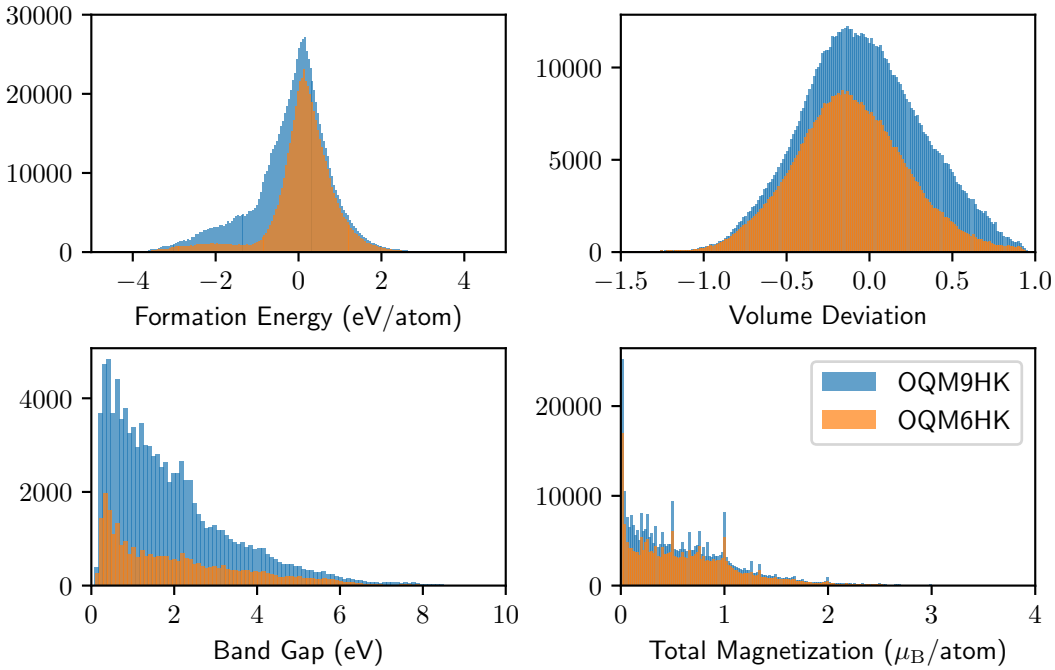

Figure 1: Histograms of formation energy, volume deviation, band gap, and total magnetization in the OQM6HK (orange) and OQM9HK (blue) dataset, whose bin widths are 0.1 eV/atom, 0.0125, 0.1 eV, and 0.02 μ_B/atom, respectively. No values of band gap and total magnetization less than 0.01 eV and μ_B/atom, respectively, are included in the histograms.

are 36%, 57%, 163%, and 36% of the corresponding database MAE, respectively. Thus, only for band gap the leading error comes from the prediction, and for other cases from the database. The worse predictions for band gaps of insulating materials are due to the high occupancy (95%) of metal materials in the OQM6HK dataset.

In the next section, we introduce our new graph dataset describing its detailed construction and explaining features of new graphs by examples. We present experimental results in §3 and discuss significance of this graph dataset in §4. In this paper, crystal graph generative models are not discussed at all, although it can be considered as one of main components in the data mining system.

Table 1: The OQM6HK and OQM9HK datasets.

	OQM6HK	OQM9HK
Materials	561,888	881,678
- Training	449,867	705,447
- Validation	56,289	87,618
- Testing	55,732	88,613
Formulas	338,135	498,146
- Training	270,527	398,536
- Validation	33,804	49,805
- Testing	33,804	49,805
Metals	531,520	779,354
	94.6%	88.4%
Insulators	30,368	102,324
	5.4%	11.6%
Nonmagnets	307,039	529,779
	54.6%	60.1%
Magnets	254,849	351,899
	45.4%	39.9%

2 The OQM9HK Graph Dataset

There are 913,045 distinct entries with the formation energy less than 5 eV in the OQMD v1.5. Although the entry's properties are calculated with the static configuration at the optimized structure obtained by the relaxation calculation, 742 entries of them have the failed relaxation, and therefore are excluded from the dataset. We regard as an abnormal entry those having at least one huge stress component of the relaxation ($>$ 10 bar) or a large volume mismatch between the static and relaxed structure ($>$ 1 Å^3/atom). We also found a few obviously abnormal entries. All the abnormal entries are excluded from the dataset. 22,958 entries completely lack records of the relaxation, but we consider that almost all of them should be reliable, confirming the equivalence or similarity between entries of the OQMD v1.2 and v1.5. All the unreliable entries are excluded from the dataset. The PyMatGen's structure matcher found 21,154 duplicates in the remaining data.[16] All the duplicate entries are excluded from the dataset. Finally, the OQM9HK dataset is constructed from the 881,678 unique and normal or reliable entries.

As shown in Table 1, the number of materials in the OQM9HK dataset increases by 57% compared with one in the OQM6HK dataset, while the number of reduced chemical formulas (or chemical compositions) increases by 47%. The increase ratios of the material and formula count differ by 10 points. Thus, there is an increase in the average number of crystal polymorphs per chemical composition.

The number of atoms per unit cell of each material is from 1 to 368 for the OQM6HK dataset. Its maximum decreases to 312 for the OQM9HK dataset due to improving the reliability of entries, although OQMD v1.5's one is 368.

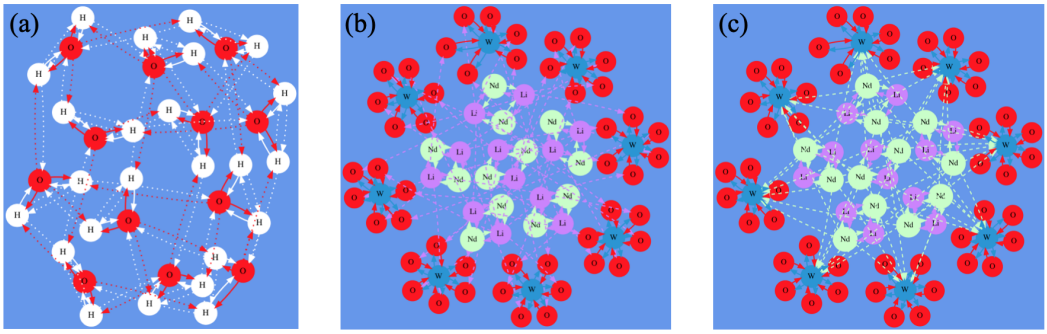

Figure 2: The crystal graphs of (a) the ice I_h and (b, c) the $Li_3Nd_3(WO_6)_2$ garnet. The solid and dotted or dashed arrows represent 1NN and 2NN directed edges, respectively. The dashed arrows depict 2NN edges from Li to O node in the panel (b), and from Nd to W node in the panel (c). There are also 2NN directed edges opposite to those depicted in the panels (b) and (c).

Its median is 4 for both datasets, and its mean is 5.33 and 6.82 for the OQM6HK and OQM9HK dataset, respectively. Thus, there is an increase in the average number of atoms per unit cell. The atomic number is limited to a number from 1 to 83, or from 89 to 94 in the OQMD, and thus there are 89 chemical elements in both datasets. As a side note, the 118 chemical elements have been identified as of 2022. Thus, neither dataset includes 29 of radioactive elements that have already been identified. The number of chemical elements contained in each material is a number from 1 to 7 for the OQM6HK dataset. Its maximum increases to 10 for the OQM9HK dataset. Its median is 3 for both datasets. The ternary materials occupy 69% and 60% of the total for the OQM6HK and OQM9HK dataset, respectively. Thus, there is a decrease in the occupancy of ternary materials.

The histogram of formation energy in the OQM9HK dataset, as shown in the top left panel of Fig. 1, differs so much from one of the OQM6HK dataset. There is a large increase in the range below 0 eV/atom, and no longer the valley as the OQM6HK dataset has around -1 eV/atom.

As in the CGNN paper,[15] instead of the unit cell volume we employ as the target property the volume deviation defined as

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

where V_a denotes the total atomic volume, and V_c the unit cell volume. The volume deviation is less than unity by this definition, but the unit cell volume less than a half of the atomic volume gives the volume deviation less than negative unity. The histogram of volume deviation (the top right panel in Fig. 1) shows that the OQM9HK dataset contains more entries with relatively large volume per atom than the OQM6HK dataset.

We regard as metals materials with the band gap less than 0.01 eV, otherwise as insulators. The number of insulating materials is 102k (12% of the total), which are greater than one of the OQM6HK dataset, 30k (5% of the total). The histogram of band gap (the bottom left panel of Fig. 1) increases entirely, but the increase ratio below 1.5 eV exceeds one in the other range. The mean and standard deviation of the whole data are 0.229449 and 0.841111 eV, respectively, which are later used for calculating z-scores to evaluate insulating probabilities.

We regard as nonmagnets materials with the total magnetization less than 0.01 μ_B/atom, otherwise as magnets. The number of magnetic materials is 352k (40% of the total), which are greater than one of the OQM6HK dataset, 255k (45% of the total). The occupancy of magnetic materials however is less than one of the OQM6HK dataset. The histogram of total magnetization (the bottom right panel of Fig. 1) increases especially below about 1 μ_B/atom. The mean and standard deviation of the whole data are 0.259047 and 0.486081 μ_B/atom, respectively, which are later used for calculating z-scores to evaluate magnetic probabilities.

We use the same algorithm to construct crystal graphs as described in the CGNN paper except for the extension described below.[15] The original graph construction uses the 2-MEANS clustering method to extract the nearest cluster of neighbors. New one uses the 3-MEANS clustering method to extract the first and second-nearest clusters of neighbors. We call graphs created by the 2-MEANS and 3-MEANS method, respectively, NC2 and NC3. These names stand for the number of cluster centers. A binary edge-labeled (BEL) graph is created from NC2 and NC3 graphs by the following process. (1) All the NC2 edges are labeled as 1NN, and (2) all the NC3 edges exceeding all the 1NN edges are labeled as 2NN. Note that this processing is necessary because the first-nearest cluster of the NC2 graph may differ from one of the NC3 graph.

For example, the crystal graph of the ice I_h is shown in Fig. 2(a). If the 2NN edges have disappeared, the crystal graph would be represented as an ensemble of unconnected sub-graphs representing individual H_2O molecules. In this case, the 2NN edges between H and O nodes apparently play a role of hydrogen bonds in chemistry. Another example is the $Li_3Nd_3(WO_6)_2$ garnet, whose crystal graph is shown in Fig. 2(b) and (c). If the 2NN edges have disappeared, the Li_2Nd_2 sub-graph would be unconnected to the WO_6 sub-graph. Therefore, one can consider 2NN edges indispensable to make more materials representable in crystal graph.

3 Experiments

3.1 Experimental Setup

We use CGNN models to understand how graph neural networks work on the OQM9HK dataset. The CGNN architec-

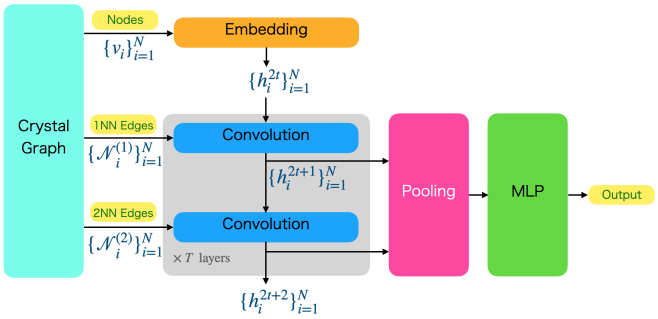

Figure 3: The CGNN architecture for binary edge-labeled graphs. Every node v_i has the 1NN and 2NN neighbor sets, $\mathcal{N}_i^{(1)}$ and $\mathcal{N}_i^{(2)}$. The embedding layer creates the embedding vector h_i^0 for each node v_i. The stacking layers of the 1NN and 2NN convolution block from $t = 0$ to $t = T$ are sequentially connected. All the hidden states are fed into the pooling layer. The MLP predicts the graph property from the graph-level hidden state made by the pooling.

Table 2: Formation energy prediction MAEs are shown in eV/atom. The best MAEs for single and ensemble models are represented in blue and red, respectively. The corresponding database MAE is 0.0848 eV/atom.[15]

Graph	d_h	Single	Ensemble
NC2	96	$0.05310 \pm 6.6 \times 10^{-5}$	0.04540
	192	$0.04926 \pm 2.2 \times 10^{-4}$	0.04264
NC3	96	$0.04850 \pm 2.2 \times 10^{-4}$	0.04108
	192	$0.04561 \pm 3.4 \times 10^{-4}$	0.03898
BEL	96	$0.04439 \pm 3.8 \times 10^{-4}$	0.03787
	192	$0.04249 \pm 3.7 \times 10^{-4}$	0.03658
Ens.	96	$0.03903 \pm 3.5 \times 10^{-5}$	0.03583
	192	$0.03712 \pm 1.3 \times 10^{-4}$	0.03433

ture consists of an embedding layer, convolution blocks, a gated pooling, and a multilayer perceptron (MLP). The convolution block has an edge-wise network that products two hidden states of the edge's ends, namely EdgeNet. For the details of this architecture we refer to the CGNN paper.[15] As shown in Fig. 3, the CGNN architecture is extended to be applied to the BEL graphs. The original CGNN has only four convolution blocks ($T = 4$), while the extended one has four additional convolution blocks introduced by 2NN edges. The softplus activation is applied to the output of the CGNN model for non-negative targets, that is, the band gap and total magnetization.

We employ the numerical library PyTorch v1.10 to train CGNN models on this dataset,[‡] and use an Nvidia T4 or P100 GPU for GPU-accelerated computing. The hyperparameters are almost the same as in the CGNN paper. The model is trained for 300 epochs by the ADAM optimizer with the batch size of 512 and the weight decay of 1×10^{-6} in a decoupled manner.[17] We use the cosine annealing method for the learning rate decay.[18] The learning rate is initially set to 1×10^{-3}, and decayed to its minimum of 1×10^{-4}. We use the learning rate warmup dedicated to the ADAM optimizer, namely the untuned warmup.[19] Really, to linearly warmup the learning rate, the learning rate is multiplied by the dampening factor during the warmup period of 2,000 steps. We employ the mean squared error as the training loss, and use the MAE as the evaluation metric. The MAE scores mainly interest us, but the root mean squared error (RMSE) scores are presented in Appendix.

Every selected CGNN model is trained once for each of three random seeds fixed through the experiments. The score of the single model is calculated as the mean of three sample scores. The three samples are members of the ensemble model for every configuration. We call it trio-ensemble in this paper. We also create an ensemble model composed of three models trained on the NC2, NC3, and BEL graph datasets, respectively. This graph-ensemble model is regarded as a single model in tables that show MAE values. The full ensemble model is the ensemble of 9 models created by collecting 3 models for each graph dataset.

The insulating and magnetic probabilities of a material are calculated as

$$p = \sigma(\zeta_{\text{train}}(y_{\text{test}})), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, y_{test} a predicted value of either band gap or total magnetization, and $\zeta_{\text{train}}(\cdot)$ 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.

3.2 Formation Energy

We use complete CGNN models for formation energy predictions. As shown in Table 2, the complete CGNN model with the hidden dimension (d_h) of 96 trained on the NC2 graph dataset gives an MAE of 53 meV/atom lower than the database MAE (85 meV/atom). The MAE score decreases by 5 and 9 meV/atom when this model is trained on the NC3 and BEL graph dataset, respectively. Upon increasing d_h to 192, the MAE score for the NC2 graph dataset improves to 49 meV/atom, while the MAE decrease changes to 4 and 7 meV/atom for the NC3 and BEL graph dataset, respectively. The best single model score of 42.5 meV/atom is given by the model with d_h of 192 trained on the BEL graph dataset. The trio-ensemble model for the best configuration gives an MAE of 36.6 meV/atom, while the graph-ensemble model gives an MAE of 37.1 meV/atom. The full ensemble model gives the best score of 34.3 meV/atom, which is 40% of the corresponding database MAE.

[‡]Our source code becomes publicly available at CGNN v1.1 (https://github.com/Tony-Y/cgnn).

Table 3: Volume deviation prediction MAEs. The best MAEs for single and ensemble models are represented in blue and red, respectively. The corresponding database MAE is 0.0270.[15]

Graph	d_h	Single	Ensemble
NC2	192	$0.01843 \pm 7.6 \times 10^{-5}$	0.01589
	288	$0.01795 \pm 1.2 \times 10^{-4}$	0.01558
NC3	192	$0.01690 \pm 1.6 \times 10^{-5}$	0.01450
	288	$0.01644 \pm 1.4 \times 10^{-4}$	0.01424
BEL	192	$0.01563 \pm 1.9 \times 10^{-4}$	0.01363
	288	$0.01496 \pm 1.3 \times 10^{-4}$	0.01331
Ens.	192	$0.01383 \pm 4.5 \times 10^{-5}$	0.01281
	288	$0.01354 \pm 4.4 \times 10^{-5}$	0.01263

Table 4: Metal-insulator classification ROC-AUCs. The best AUCs for single and ensemble models are represented in blue and red, respectively. The corresponding database AUC is 0.9564.[15]

Graph	Single	Ensemble
NC2	$0.95533 \pm 4.1 \times 10^{-3}$	0.96490
NC3	$0.95158 \pm 5.8 \times 10^{-3}$	0.96159
BEL	$0.96449 \pm 9.1 \times 10^{-4}$	0.97127
Ens.	$0.97026 \pm 1.1 \times 10^{-3}$	0.97338

3.3 Unit Cell Volume

We use noEdgeNet CGNN models for volume deviation predictions. As shown in Table 3, the noEdgeNet CGNN model with d_h of 192 trained on the NC2 graph dataset gives an MAE of 1.84×10^{-2} lower than the database MAE (2.70×10^{-2}). The MAE score decreases by 0.15×10^{-2} and 0.28×10^{-2} when this model is trained on the NC3 and BEL graph dataset, respectively. Upon increasing d_h to 288, the MAE score for the NC2 graph dataset improves to 1.80×10^{-2}, while the MAE decrease changes to 0.30×10^{-2} for the BEL graph dataset, but is almost the same for the NC3 graph dataset. The best single model is obtained with d_h of 288 for the BEL graph dataset, which gives an MAE of 1.50×10^{-2}. The trio-ensemble model for the best configuration gives an MAE of 1.33×10^{-2}, while the graph-ensemble model gives an MAE of 1.35×10^{-2}. The full ensemble model gives the best score of 1.26×10^{-2}, which is 47% of the corresponding database MAE.

3.4 Band Gap

We use complete CGNN models with d_h of 192 for band gap predictions. The model trained on the NC2 graph dataset gives an MAE of 82.4 meV as shown in Table 5 (left). The MAE score decreases by 8.4 and 7.1 meV when this model is trained on the NC3 and BEL graph dataset, respectively. The best single model is obtained for the NC3 graph dataset. The trio-ensemble model for the best configuration gives an MAE of 68.2 meV, while the graph-ensemble model gives an MAE of 69.1 meV. The full ensemble model gives the best score of 66.7 meV.

For evaluation on the metal subset, as shown in Table 5 (middle), the MAE scores (∼20 meV) are lower than the database MAE (46.1 meV). The best MAE of 19.5 meV is given by both single and trio-ensemble model for the NC3 graph dataset.

For evaluation on the insulator subset, as shown in Table 5 (right), the best single model becomes one trained on the BEL graph dataset and its MAE is 0.489 eV. The best MAE of 0.418 eV given by the full ensemble model is lower than our previously obtained value on the OQM6HK dataset, but is 122% of the corresponding database MAE (0.341 eV). Therefore, the prediction uncertainty is still higher than the database uncertainty.

The metal-insulator classification ROC-AUC of the model trained on the BEL graph dataset is 96.4% as shown in Table 4, which is higher than the database AUC (95.6%). The trio-ensemble model for the best configuration gives an AUC of 97.1%, while the graph-ensemble model gives an AUC of 97.0%. The full ensemble model gives the best AUC of 97.3%.

3.5 Total Magnetization

We use noEdgeNet CGNN models with d_h of 288 for total magnetization predictions. The model trained on the NC2 graph dataset gives an MAE of 60.9 mμ_B/atom as shown in Table 6 (left). The MAE score decreases by 2.7 and 3.5 mμ_B/atom when this model is trained on the NC3 and BEL graph dataset, respectively. The best single model is obtained for the BEL graph dataset. The trio-ensemble model for the best configuration gives an MAE of 53.2 mμ_B/atom, while the graph-ensemble model gives an MAE of 52.7 mμ_B/atom. The full ensemble model gives the best score of 50.7 mμ_B/atom.

For evaluation on the nonmagnet subset, as shown in Ta-

Table 5: Band gap prediction MAEs evaluated on (left) the whole dataset, (middle) the metal subset, and (right) the insulator subset are shown in eV. The best MAEs for single and ensemble models are represented in blue and red, respectively. The corresponding database MAE is 0.1806 eV for the whole data, 0.0461 eV for the metal subset, and 0.3412 eV for the insulator subset.[15]

	Whole		Metal		Insulator	
Graph	Single	Ensemble	Single	Ensemble	Single	Ensemble
NC2	$0.08244 \pm 6.0 \times 10^{-4}$	0.07663	$0.02230 \pm 1.4 \times 10^{-3}$	0.02230	$0.54628 \pm 7.5 \times 10^{-3}$	0.49566
NC3	$0.07405 \pm 1.4 \times 10^{-4}$	0.06825	$0.01952 \pm 8.4 \times 10^{-4}$	0.01952	$0.49463 \pm 6.0 \times 10^{-3}$	0.44404
BEL	$0.07538 \pm 2.0 \times 10^{-4}$	0.06920	$0.02173 \pm 2.1 \times 10^{-4}$	0.02173	$0.48909 \pm 2.1 \times 10^{-3}$	0.43530
Ens.	$0.06913 \pm 6.7 \times 10^{-5}$	0.06668	$0.02118 \pm 3.7 \times 10^{-4}$	0.02118	$0.43886 \pm 3.2 \times 10^{-3}$	0.41753

Table 6: Total magnetization prediction MAEs evaluated on (left) the whole dataset, (middle) the nonmagnet subset, and (right) the magnet subset are shown in μ_B/atom. The best MAEs for single and ensemble models are represented in blue and red, respectively. The corresponding database MAE is 0.0938 μ_B/atom for the whole data, 0.0211 μ_B/atom for the nonmagnet subset, and 0.3274 μ_B/atom for the magnet subset.[15]

	Whole		Nonmagnet		Magnet	
Graph	Single	Ensemble	Single	Ensemble	Single	Ensemble
NC2	$0.06087 \pm 2.8 \times 10^{-4}$	0.05553	$0.01769 \pm 1.9 \times 10^{-4}$	0.01752	$0.12574 \pm 6.8 \times 10^{-4}$	0.11262
NC3	$0.05820 \pm 3.0 \times 10^{-4}$	0.05331	$0.01684 \pm 2.4 \times 10^{-4}$	0.01667	$0.12032 \pm 3.8 \times 10^{-4}$	0.10835
BEL	$0.05735 \pm 3.7 \times 10^{-4}$	0.05322	$0.01722 \pm 7.1 \times 10^{-5}$	0.01706	$0.11762 \pm 9.8 \times 10^{-4}$	0.10754
Ens.	$0.05272 \pm 3.0 \times 10^{-5}$	0.05073	$0.01706 \pm 2.9 \times 10^{-5}$	0.01696	$0.10629 \pm 1.1 \times 10^{-4}$	0.10147

ble 6 (middle), the model trained on the NC2 graph dataset gives an MAE of 17.7 mμ_B/atom lower than the database MAE (21.1 mμ_B/atom). The best single and ensemble MAE of 16.8 and 16.7 mμ_B/atom, respectively, are given by the single and trio-ensemble model for the NC3 graph dataset.

For evaluation on the magnet subset, as shown in Table 6 (right), the model trained on the NC2 graph dataset gives an MAE of 0.126 μ_B/atom lower than the database MAE (0.327 μ_B/atom). The best single model becomes one trained on the BEL graph dataset and its MAE is 0.118 μ_B/atom. The trio-ensemble model for the best configuration gives an MAE of 0.106 μ_B/atom, while the graph-ensemble model gives an MAE of 0.108 μ_B/atom. The full ensemble model gives the best score of 0.101 μ_B/atom, which is 31% of the corresponding database MAE.

The magnet classification ROC-AUC of the model trained on the NC2 graph dataset is 95.5% as shown in Table 7, which is higher than the database AUC (86.9%). The best single model is obtained for the BEL graph dataset and its AUC is 95.9%. The trio-ensemble model for the best configuration gives an AUC of 96.6%, while the graph-ensemble model gives an AUC of 96.5%. The full ensemble model gives the best AUC of 96.9%.

4 Discussions

Analyzing the error distributions of the trio-ensemble models trained on the NC2, NC3, and BEL graph dataset, we can deduce that the use of NC3 graphs instead of NC2 graphs reduces overpredictions of formation energy and volume deviation, while the use of BEL graphs reduces their underpredictions in addition to the overprediction reduction. Moreover, we can find that the use of NC3 or BEL graphs reduces underpredictions of band gap but slightly increases their overpredictions, while it reduces underpredictions of total magnetization.

Materials scientists are often interested in thermodynamically stable polymorphs. It is desirable to precisely predict the ranking in polymorphic stability. We employ the mean of Kendall's tau values to measure the ranking performance. The tau scores of the trio-ensemble models trained on the NC2, NC3, and BEL graph dataset are 0.789, 0.802, and 0.813, respectively. Thus, the use of BEL graphs slightly improves the ranking performance.

The experimental results and additional analysis show that the models trained on the BEL graph dataset outperform those trained on the NC2 or NC3 graph dataset. Hence, we can conclude that BEL graphs are preferable to NC2 or NC3 graphs as inputs of graph neural networks predicting the materials properties.

Although it contains the unit cell volume prediction task, this graph dataset does not contain enough information on crystal structures for structure prediction tasks. However, one can easily retrieve the entire structural information of each entry because each entry has the identification number of the corresponding calculation entry in the OQMD v1.5.[§] On the basis of the retrieved structural information, one can invent a crystal structure prediction task.

As shown in the experiments and discussions above, this graph dataset has the decided advantages against the previous one. Therefore, the OQM9HK graph dataset would facilitate studies on graph representation learning in materials science.

Table 7: Magnet classification ROC-AUCs. The best AUCs for single and ensemble models are represented in blue and red, respectively. The corresponding database AUC is 0.8688.[15]

Graph	Single	Ensemble
NC2	$0.95510 \pm 9.9 \times 10^{-5}$	0.96196
NC3	$0.95880 \pm 3.9 \times 10^{-4}$	0.96511
BEL	$0.95927 \pm 1.7 \times 10^{-4}$	0.96537
Ens.	$0.96583 \pm 2.6 \times 10^{-4}$	0.96857

References

[1] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 40–48. JMLR.org, 2016.

[2] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks, 2020. `https://github.com/graphdeeplearning/benchmarking-gnns`.

[§]We can see a calculation entry through the official online database instead of fetching it from an OQMD v1.5 database installed on a Linux server. For example, the first entry in this graph dataset has the calculation ID of 1299782, and then its calculation entry's URL becomes the following one: https://oqmd.org/analysis/calculation/1299782

[3] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs, 2020. https://ogb.stanford.edu.

[4] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs, 2021.

[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] 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.

[7] 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.

[8] Christopher J. Shallue and Andrew Vanderburg. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astronomical Journal*, 155(2):94, Jan 2018.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[10] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. https://alphafold.ebi.ac.uk.

[11] James Kirkpatrick, Brendan McMorrow, David H. P. Turban, Alexander L. Gaunt, James S. Spencer, Alexander G. D. G. Matthews, Annette Obika, Louis Thiry, Meire Fortunato, David Pfau, Lara Román Castellanos, Stig Petersen, Alexander W. R. Nelson, Pushmeet Kohli, Paula Mori-Sánchez, Demis Hassabis, and Aron J. Cohen. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, 2021.

[12] Amir Masoud Rahmani, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh, and Rizwan Ali Naqvi. Machine learning (ml) in medicine: Review, applications, and challenges. *Mathematics*, 9(22):2970, Nov 2021.

[13] Chunming Xu and Scott A. Jackson. Machine learning and complex biological data. *Genome Biology*, 20:76, Apr 2019.

[14] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, September 2020.

[15] Takenori Yamamoto. Crystal graph neural networks for data mining in materials science. Technical report, Research Institute for Mathematical and Computational Sciences, LLC, Yokohama, Japan, 2019. https://github.com/Tony-Y/cgnn.

[16] 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.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[18] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[19] Jerry Ma and Denis Yarats. On the adequacy of untuned warmup for adaptive optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8828–8836, May 2021.

A Additional Information

Table A.1: Formation energy prediction RMSEs are shown in eV/atom. The best RMSEs are represented in blue and red for single and ensemble models, respectively. The corresponding database RMSE is 0.1242 eV/atom.[15]

Graph	d_h	Single	Ensemble
NC2	96	$0.10812 \pm 2.5 \times 10^{-4}$	0.09866
	192	$0.10363 \pm 1.0 \times 10^{-3}$	0.09593
NC3	96	$0.10141 \pm 4.3 \times 10^{-4}$	0.09180
	192	$0.09745 \pm 3.2 \times 10^{-4}$	0.08947
BEL	96	$0.09281 \pm 8.3 \times 10^{-4}$	0.08480
	192	$0.09109 \pm 4.9 \times 10^{-4}$	0.08392
Ens.	96	$0.08563 \pm 1.1 \times 10^{-4}$	0.08214
	192	$0.08375 \pm 4.0 \times 10^{-4}$	0.08083

Table A.2: Volume deviation prediction RMSEs. The best RMSEs are represented in blue and red for single and ensemble models, respectively. The corresponding database RMSE is 0.0421.[15]

Graph	d_h	Single	Ensemble
NC2	192	$0.03576 \pm 3.2 \times 10^{-4}$	0.03261
	288	$0.03549 \pm 6.3 \times 10^{-4}$	0.03253
NC3	192	$0.03298 \pm 1.0 \times 10^{-4}$	0.03021
	288	$0.03250 \pm 3.8 \times 10^{-4}$	0.03009
BEL	192	$0.03071 \pm 3.7 \times 10^{-4}$	0.02840
	288	$0.02983 \pm 2.7 \times 10^{-4}$	0.02810
Ens.	192	$0.02844 \pm 5.2 \times 10^{-5}$	0.02739
	288	$0.02832 \pm 8.6 \times 10^{-5}$	0.02743

Table A.3: Band gap prediction RMSEs evaluated on (left) the whole dataset, (middle) the metal subset, and (right) the insulator subset are shown in eV. The best RMSEs for single and ensemble models are represented in blue and red, respectively. The corresponding database RMSE is 0.5288 eV for the whole data, 0.3568 eV for the metal subset, and 0.6794 eV for the insulator subset.[15]

	Whole		Metal		Insulator	
Graph	Single	Ensemble	Single	Ensemble	Single	Ensemble
NC2	$0.39881 \pm 3.6 \times 10^{-3}$	0.36846	$0.21713 \pm 1.4 \times 10^{-2}$	0.19286	$1.01004 \pm 3.7 \times 10^{-2}$	0.94653
NC3	$0.35293 \pm 7.5 \times 10^{-3}$	0.32223	$0.19267 \pm 9.4 \times 10^{-3}$	0.17068	$0.89354 \pm 2.4 \times 10^{-2}$	0.82458
BEL	$0.35588 \pm 5.5 \times 10^{-3}$	0.32260	$0.21228 \pm 5.0 \times 10^{-3}$	0.19001	$0.86942 \pm 1.0 \times 10^{-2}$	0.79261
Ens.	$0.31839 \pm 7.3 \times 10^{-4}$	0.30734	$0.17623 \pm 2.6 \times 10^{-3}$	0.16811	$0.80227 \pm 2.3 \times 10^{-3}$	0.77779

Table A.4: Total magnetization prediction RMSEs evaluated on (left) the whole dataset, (middle) the nonmagnet subset, and (right) the magnet subset are shown in μ_B/atom. The best RMSEs for single and ensemble models are represented in blue and red, respectively. The corresponding database RMSE is 0.4003 μ_B/atom for the whole data, 0.1399 μ_B/atom for the nonmagnet subset, and 0.7824 μ_B/atom for the magnet subset.[15]

	Whole		Nonmagnet		Magnet	
Graph	Single	Ensemble	Single	Ensemble	Single	Ensemble
NC2	$0.18581 \pm 1.8 \times 10^{-3}$	0.17451	$0.10130 \pm 2.1 \times 10^{-3}$	0.09602	$0.26640 \pm 2.2 \times 10^{-3}$	0.24970
NC3	$0.18494 \pm 9.5 \times 10^{-4}$	0.17424	$0.09841 \pm 7.9 \times 10^{-4}$	0.09318	$0.26653 \pm 1.3 \times 10^{-3}$	0.25083
BEL	$0.18328 \pm 1.9 \times 10^{-3}$	0.17406	$0.09997 \pm 1.1 \times 10^{-4}$	0.09520	$0.26275 \pm 3.3 \times 10^{-3}$	0.24938
Ens.	$0.16945 \pm 1.3 \times 10^{-3}$	0.16564	$0.09221 \pm 1.5 \times 10^{-4}$	0.09034	$0.24305 \pm 2.2 \times 10^{-3}$	0.23747