Predicting the Materials Properties Using a 3D Graph Neural Network with Invariant Representation

BOYU ZHANG1, MUSHEN ZHOU2, JIANZHONG WU2, AND FUCHANG GAO.1

1Department of Mathematics and Statistical Science, University of Idaho, Moscow, ID 83843, United States
2Department of Chemical and Environmental Engineering, University of California, Riverside, CA 92521, United States

Corresponding author: Boyu. Zhang (e-mail: boyuz@uidaho.edu).

This work is financially supported by the National Science Foundation Harnessing the Data Revolution Big Idea under Grant No. NSF 1940118 and Grant No. NSF 1940270.

ABSTRACT Accurate prediction of physical properties is critical for discovering and designing novel materials. Machine learning technologies have attracted significant attention in the materials science community for their potential for large-scale screening. Graph Convolution Neural Network (GCNN) is one of the most successful machine learning methods because of its flexibility and effectiveness in describing 3D structural data. Most existing GCNN models focus on the topological structure but overly simplify the three-dimensional geometric structure. However, in materials science, the 3D-spatial distribution of atoms is crucial for determining the atomic states and interatomic forces. This paper proposes an adaptive GCNN with a novel convolution mechanism that simultaneously models atomic interactions among all neighboring atoms in three-dimensional space. We apply the proposed model to two distinctly challenging materials properties prediction problems. The first is Henry’s constant for gas adsorption in Metal-Organic Frameworks (MOFs), which is notoriously difficult because of its high sensitivity to atomic configurations. The second is the ion conductivity in solid-state crystal materials, which is difficult because of the few labeled data available for training. The new model outperforms existing graph-based models on both data sets, suggesting that the critical three-dimensional geometric information is indeed captured.

INDEX TERMS Graph Convolution Neural Network, Metal-Organic Frameworks, Henry’s Constant, Property Prediction

I. INTRODUCTION

Accurate and rapid prediction of physical properties plays an essential role in discovering and designing novel materials. Ab initio methods [1] [2] are commonly used to predict the materials properties. However, ab initio methods, such as Density-Functional Theory (DFT) [3] and wave function theories [4], are time-consuming. An alternative method is to predict these properties using machine learning (ML) technologies [5]. By modeling the structures of materials and atomic interactions, ML algorithms are much faster than DFT calculations and yet provide similar accuracy [6].

Although ML techniques have succeeded in many fields, such as computer vision and natural language processing, successfully applying ML techniques in materials science is still significantly challenging [7]. The atomic structures of the materials are complex three-dimensional spatial structures with an unfixed number of atoms. The classic machine learning models, such as Convolutional Neural Network (CNN), are better at processing fixed-size data. In contrast, graph-based models, such as Message Passing Neural Network (MPNN) [8] and Graph Convolutional Neural Networks (GCNN) [9], have the advantage of flexibility in describing variable data. Moreover, the message passing (for MPNN) and convolution (for GCNN) could effectively stimulate the atomic interaction in the materials. Recently, graph-based models have attracted significant attention because of the flexibility and effectiveness in describing 3D structures formed by discrete particles [10]–[13].

Although the model’s effectiveness has been well demonstrated, the current GCNN has some shortcomings. GCNN describes the structures of the materials with undirected graphs, where nodes represent the atoms while edges express the relationship between bonded atoms. The graph-based
description is concise and efficient but ignores the materials three-dimensional topology, which critically impacts the properties of the materials. In addition, the GCNN updates the state of each atom by accumulating the impact of the surrounding bonded atoms. This design is based on the assumption that the surrounding atoms independently influence the current atom. However, the surrounding atoms exert their influence on the current atom simultaneously. The model would be more effective if it were true to reality [14].

This paper presents the k-Nearest Atoms Graph Convolution Network (k-NAGCN) for predicting the physical properties of materials. The proposed model improves the accuracy of attribute prediction by introducing three-dimensional topological structure information. We also designed an innovative convolutional structure to simultaneously calculate the influence of surrounding atoms on the current atom. Besides, we developed an algorithm to calculate translate and rotation invariant descriptors for the local atomic distribution. We tested the new model with two distinctly challenging materials properties prediction problems. The first is Henry’s constant for gas adsorption in Metal-Organic Frameworks (MOFs), which is notoriously difficult because the results are highly sensitive to inter-atomic separations. The second example is the prediction of ion conductivity in solid-state crystal materials. The task is also difficult due to the lack of labeled data available for training. The k-NAGCN model presents superior performance on both datasets, suggesting that it indeed captures important three-dimensional geometric information. The contributions of this paper are as follows.

- Proposed an end-to-end system for predicting the physical properties of the materials.
- Modeled the three-dimensional materials structures with a graph-based model and proposed a novel convolution mechanism that better simulates physical laws.
- Developed a rotation- and translation-invariant descriptor to model the local topology of atoms.

II. RELATED WORK

The emergence of large-scale datasets of materials [15]–[17] greatly empowers the applications of novel machine learning technologies in the field of materials science [18]. The representation of materials plays a critical role in building an effective machine learning system, and it is challenging due to the complexity of the chemical structures. One strategy is to describe the materials using a group of physical properties. Cubuk et al. [19] trained a support vector machine (SVM) based on thirty features and then used transfer learning to get a formula-based predictor. The formula-based predictor offered significant advantages in efficiency, being able to screen billions of materials in a short period. Allam et al. [10] employed DFT modeling approach to predict basic quantum mechanical quantities. The predicted features were combined with the structure feature to obtain a neural network to predict molecular electrode materials. Feature-based methods rely on expertise in materials science, and their performance also depends on the effectiveness of the features used.

With large-scale data available, researchers in recent years tend to use data-driven approaches that make greater use of low-level structural information. Hoffmann et al. [20], for instance, constructed a smooth and continuous three-dimensional density representation of each crystal based on different atomic positions and designed an autoencoder/decoder network to predict innovative materials. The method benefits from the detailed structural information but requires high computational power. Zheng et al. [21] presented the atomic configuration as a two-dimensional periodic table and normalized the table to mimic a digital image. They trained a multitask Convolutional Neural Networks (CNN) model that outputs the lattice parameters and enthalpy of formation simultaneously. The experimental results supported that their method achieved competitive performance on ICSD [15] and OQMD [17]. Schütt et al. [12] also described materials using periodic tables. The proposed architecture, Schnet, used a continuous filter convolutional layers to model local correlations and yielded excellent performance for predicting a group of molecular properties.

Graph-based models could handle variant inputs and thus have attracted much attention in chemistry. Xie et al. [13] proposed Crystal Graph Convolutional Neural Network (CGCNN) that described crystals using undirected graphs, where the vertices and edges corresponded to the atoms and bonds, respectively. CGCNN captured the atomic interactions using the graph convolution layers, and a pooling layer embedded the graph into the feature space. The CGCNN model showed excellent performance on various properties, including formation energy, bandgap, Fermi energy, and elastic properties. Sanyal et al. [22] adapted the original CGCNN using multitask learning [23], and their experiments showed that sharing parameters in the lower level improved the prediction performance.

Domain knowledge can be integrated into data-driven models to improve performance. Chen et al. [24] developed a universal model to predict the physical properties of both molecules and crystals by incorporating state variables, such as temperature, pressure, and entropy. Lee and Asahi [25] used transfer learning, pretrained a GCNN model with big data, and then used it for predicting target properties with relatively small data. Cho et al. [26] added a relative position matrix into the model and provided a mechanism to handle the feature update process. The proposed model, 3DGDCN, showed good performance for predicting molecular properties and biochemical activities.

III. METHOD

This section explains the details of the proposed k-NAGCN. The model accepts the structures of the materials as input. After modeling the structure, convolution operations are performed on the graph, and finally, the features of each node are aggregated, and the target attributes are estimated. The following subsections discuss graph-based structure representa-
tion, convolution operations, and the algorithm that produces translation- and rotation-invariant structural descriptors in three-dimensional space.

A. GRAPH-BASED REPRESENTATION
To capture the three-dimensional topology of each material, we used the following graph-based representations: A material containing \( N \) atoms was represented by an undirected graph \( G = (V, E) \), where the vertices \( V = \{v_1, \ldots, v_N\} \) represent the atoms, and the edges \( E = \{e_{i,j}|i, j \in \{1, \ldots, N\}, i \neq j\} \) represent the connections between atoms (not necessarily bonds, as some atoms in the model may not be chemically bonded.) Each vertex in the graph stores two pieces of information: atomic feature of the atom and the spatial location of the atom \( X \). Each edge stores one edge feature. (See Figure 1(a)).

We chose a hyperparameter \( M \). For each type of element in the dataset, we used an \( M \)-dimensional feature vector to store the prior chemistry knowledge about the element. Such domain knowledge about the elements is independent of the structures of the materials. These vectors were calculated using the adaptive ‘atom2vector’ method [27]. With \( L \) elements in the datasets, these feature vectors form an \( M \times L \) matrix, denoted by \( F \). If the atom at vertex \( v_i \) of the graph is an element of type \( j \), the \( j \)-th row vector of the matrix \( F \) is used as the feature vector of vertex \( v_i \). Technically, this was done by using one-hot encoding [28] to represent the element type. That is, to construct an \( N \times L \) matrix \( A = (a_{ij})_{N \times L} \), such that \( a_{ij} = 1 \) if the \( i \)-th atom in the input structure is of element type \( j \), and 0 otherwise. Then, the \( i \)-th row vector of the product matrix \( AF \) gives the atomic feature of \( i \)-th vertex.

The spatial distribution of the atoms in the materials was represented by the \( N \times 3 \) coordinate matrix \( X \), in which the \( i \)-th row is the atomic coordinates of \( i \)-th atom relative to the unit cell of the materials.

In most graph-based representations, the adjacency matrix is used to describe the chemical bonds. This model replaced the adjacency matrix with the distance matrix because k-NAGCN calculates the graph convolution over each atom and its \( k \) nearest neighbors, whether the atoms are chemically bonded or not. This distance matrix is represented by the \( N \times N \) matrix \( E = (\|e_{i,j}\|)_{N \times N} \), where \( \|e_{i,j}\| \) is the length of the edge \( e_{i,j} \) or the distance between the \( i \)-th atom and the \( j \)-th atom in the materials.

B. K-NEAREST NEIGHBOR CONVOLUTION
The graph convolution has the most significant impact on the performance of graph convolution neural networks. This subsection explains the proposed convolution mechanism that involves the nearest neighboring atoms simultaneously.

With the materials represented by graph \( G = (V, E) \), each vertex \( v_i \) is associated with the atomic feature \( f_i \) and coordinates \( x_i \), and each edge \( e_{i,j} \) is associated with the edge feature \( e_{i,j} \), which is defined as the vector

\[
(\alpha_1\|e_{i,j}\|, \alpha_2\|e_{i,j}\|, \ldots, \alpha_{32}\|e_{i,j}\|),
\]

where \( \alpha_1, \alpha_2, \ldots, \alpha_{32} \) are some pre-chosen numbers. We made some minor effort to distinguish the edge \( e_{ij} \) and its feature vector \( e_{i,j} \).

The atomic feature \( f_i \) associated with the \( i \)-th atom is updated according to equation (1).

\[
\hat{f}_i = f_i + \text{Conv}(f_i + L_{ed}(x_{i}^{\text{out}})_{ij} \in K_i \{e_{i,j}, f_j + L_{ed}(x_{j}^{\text{out}})\})
\]

where \( \text{Conv}(\cdot) \) is a combination of the linear transformation and nonlinear activation function, \( K_i \) is the index set of the \( k \) nearest vertices of the \( i \)-th vertex, and \( x_{i}^{\text{out}} \) is the transformed coordinate calculated using Algorithm 1. After the coordinate transformation, the \( x_{i}^{\text{out}} \) goes through a linear embedding layer, and the output is added to the original atomic feature \( f_i \).

The hyper-parameter \( k \) determines how many neighbor atoms should be involved in the convolution. The proposed convolution considers the impacts of \( k \) nearest neighbor atoms simultaneously and thus, is a better simulation of real-world atomic interactions. Figure 1(b) presents the process how convolution is performed in k-NAGCN when \( k = 2 \).

C. ROTATIONAL INVARIANT LOCAL DESCRIPTOR
The convolution proposed in subsection III-B is sensitive to the order of atoms and translation. In this subsection, we propose translation and rotation invariant descriptors for the local atomic distribution.

It is vital to keep the atom sequence identical when the local structure rotates or translates. Assume \( k = 2 \) and denote the two neighboring vertices of the vertex \( v_i \) by \( v_j \) and \( v_k \). The input of \( \text{Conv}(\cdot) \) is the concatenated vector \( (f_i, X_i, e_{i,j}, f_j, X_j, e_{i,k}, f_k, X_k) \). If we rotate the geometric structure composed of \( v_i, v_j, \) and \( v_k \) along the axis that passes through \( v_i \) and the middle point of \( v_j \) and \( v_k \), until \( v_j \) and \( v_k \) switch places, the chemical feature of the atom \( v_i \) does not change after rotation, but the input of \( \text{Conv}(\cdot) \) is changed to a different vector \( (f_i, X_i, f_{i,k}, f_k, X_k', f_{i,j}, f_j, X_j') \).

Inspired by the Scale-Invariant Feature Transform (SIFT) [29], which is widely used in the field of computer vision, we introduced the following algorithm to generate input sequences for \( \text{Conv}(\cdot) \) that are invariant under translation and rotation (Algorithm 1).

For each atom \( v_i \) (\( i \in \{1, N\} \)) in the structure, we calculated the distance \( 1 \) between \( v_i \) and the rest atoms \( \{v_j\} \), where \( i \neq j \), and \( N \) is the total number of atoms in the structure.

Then, we sorted all \( v_j \) according to the distance and found the \( k \) nearest atoms \( 2 \), where \( k \) is a hyper-parameter. Denote by \( x_i \) the position of atom \( v_i \), and by \( x_{i}^{1} \) and \( x_{i}^{2} \) the positions of the closest and the second closest atoms to \( x_i \). We first find transformation \( T \) that translates \( x_i \) to the origin. Next, we calculate a matrix \( R_1 \) that rotates \( x_{i}^{1} \) to positive z-axis according to Rodrigues’ rotation formula [31]. Then, we calculated the matrix \( R_2 \) that rotate \( x_{i}^{2} \) to the yz-plane, keeping \( x_{i}^{1} \) fixed. Finally, we calculated the new coordinate

\[1\] The calculation of distance was based on pymatgen [30].
\[2\] More details about the sorting algorithm could be found at https://docs.python.org/3/library/functions.html.
FIGURE 1: (a) The input material (left) is encoded into three pieces of information, the atomic features \( F \), the edge features \( E \), and the spatial locations \( X \). (b) Based on the representation in (a), the convolution is calculated as follows: For each vertex \( v_i \), two closest neighboring atoms \( v_j \) and \( v_s \) are selected. The coordinates of their spatial locations \( x_i, x_j, x_s \) are transformed by an embedding layer and then added to the atomic features \( f_i, f_j, f_s \). Then the updated atomic features are concatenated with the edge feature \( e_{i,j}, e_{i,s} \) as the input of the convolution. The output of graph convolution \( f'_i \) is added to the original feature \( f_i \).

Algorithm 1: The rotational invariant description for local structure

**Input:** Graph \( G = \{V, E\} \), coordinate matrix \( X \), number of neighbor atoms \( k \)

**Output:** Transformed coordinate matrix \( X_{out} \)

**LOOP Process**

1. for \( i \leftarrow 1 \) to \( N \) do
2. Calculate distance between vertex \( v_i \) and \( v_j \), where \( v_j \in V, j \neq i \)
3. Sort \( v_j \) based on distance and get \( k \) nearest neighbors
4. Transform the origin of coordinate system to \( x_i \)
5. Calculate transform matrix \( R_1 \) that rotates \( x_i \) to \( [0, 0, \|x_i\|^2] \) using Rodrigues’ rotation formula
6. Update the coordinate by \( X_{R_1} = R_1 \times X_i \)
7. Calculate transform matrix \( R_2 \) that rotate \( x^2 \) to \( yz \)-plane
8. Update the coordinate by \( X_{out} = R_2 \times X_{R_1} \)
9. end for
10. return \( X_{out} \)

\( R_2 \circ R_1 \circ T(x) \) for every atom in the \( k \) nearest neighborhood. It is readily seen that this distance-based sort and rotation guarantee the rotational invariance.

IV. EXPERIMENT

A. EXPERIMENTAL DATA

The proposed k-NAGCN was tested on two experimental datasets. The first dataset, CoRE2019, contains 12763 MOFs, and the property of interest is Henry’s constant [32]. Appendix A presents the details of the calculation of Henry’s constant.

The proposed algorithm was also tested with a subset of the ICSD database built for ion conductivity prediction. The properties of interest of the ICSD data include minimum bond length, large Li site, and percolation radius. In detail, there are 1758 crystals available for the bond length prediction, 2818 crystals available for percolation radius prediction, and 894 crystals available for Li site prediction. These properties were calculated through topological analysis and could be used as indicators for high potential conductors. Readers could refer to [33] for more details on the calculation of these properties.

B. IMPLEMENTATION DETAILS

1) Data cleaning for CoRE2019.

The original CoRE2019 MOF database contains 12763 MOF structures, and the range of the Henry’s constant covers more than 30 orders of magnitude. We examined the structure files and found the broad range of Henry’s constant was caused by unreasonable structures. In a small portion of MOFs, atoms were unreasonably close to each other, and the overlapped atoms led to high attraction and high Henry’s constant. Meanwhile, there were another group of MOFs
that had tightened confined geometry, and the thin structures resulted in extremely low Henry’s constant (at the order of \(10^{-10}\)).

We got rid of MOFs with atomic distance smaller than 0.8 Å to avoid overlapped atoms. We also filtered out the MOFs with extremely low Henry’s constant by removing structures with a surface area less than 50 Å² (meaning that there will be very little space and very unlikely for surface adsorption to happen). The cleaned dataset contains 10136 MOFs. The details of the above sample selection process can be found in Appendix B.

2) Value transformation of Henry’s Constant

The Henry’s constants still covered several orders of magnitudes. During the backpropagation of the gradient, the tremendous values would cause gradient explosions and make the training infeasible. Existing research ( [34]) predicted the value \(\log_{10}(H_{cc})\) instead of the widely varied original value.

In this paper, we transformed the original value to \(\log_{10}(H_{cc})\) and \(\log_{10}(H_{cc} + 1)\). The two terms were close with large \(H_{cc}\), but the latter term put less weight on materials with low Henry’s constants. The rationale of introducing the latter term \(\log_{10}(H_{cc} + 1)\) is that materials with high Henry’s constant were favored for gas adsorption, while materials with extremely low Henry’s constant was not of the practical interest.

3) Experimental setup and parameter selection

10-folded cross-validation was used throughout the experiments to ensure the fairness of the evaluation. In detail, samples were randomly divided into training set 80%, validation set 10%, and testing set 10%. The network was trained using the training set, and the performance was evaluated on the testing set. The validation set was used to select hyperparameters during the training process, such as numbers of training epochs and convolution layers. The process was repeated ten times, and the performance was evaluated using the average of the Mean Absolute Error (MAE).

The following parameters were shared for all tasks. The batch size is set as 32 to avoid overfitting. Adam is used as the optimizer to find optimized parameters. The initial learning rate was set as 0.0003. The learning rate was reduced by a factor of 0.1 for every 50 epoch.

The experiments were performed on a Microsoft Azure virtual machine with 24 vCPU (2.5GHz), 112 GB ram, and 1 V100 GPU (5120 CUDA Cores and 32GB ram).

C. RESULTS

We first predicted \(H_{cc}\) of the CoRE2019. All 10136 MOFs were used in the experiments. Both \(\log_{10}(H_{cc})\) and \(\log_{10}(H_{cc} + 1)\) were predicted. The performance was compared with two graph-based models, CGCNN [13] and 3DGCN [26], which were designed for materials properties prediction. The MT-CGCNN [22] wasn’t involved in the comparison because our Core2019 and ICSD datasets didn’t include multiple labels. The MT-CGCNN without multiple labels was equivalent to CGCNN.

Both CGCNN and 3DGCN were implemented according to the original literature and adapted to fit the CoRE2019 data. For CGCNN, the network structure includes one feature embedding layer, three graph convolution layers, and three fully-connected layers based on the performance of the validation data. For the 3DGCN, max-pooling was used because the original literature reported max-pooling outperformed average pooling. Our experiment revealed that 3DGCN with max-pooling outperformed the 3DGCN with average pooling on Core2019 and ICSD datasets. The k-NAGCN contained three graph convolution layers and two fully-connected layers, and average pooling was used. 10-folded cross-validation was used for all three models. The performance was evaluated using three different metrics, Mean Absolute Error, Mean Absolute Percentage Error, and \(R^2\) score. The results are summarized in Table 1.

\(k\)-NAGCN outperformed CGCNN and 3DGCN on \(H_{cc}\) prediction of CoRE2019, with the lowest MAE, MAPE, and highest \(R^2\) score. This result proved the two hypotheses we made in the previous section. First, 3DGCN and \(k\)-NAGCN achieved better results in \(H_{cc}\) prediction compared with CGCNN, proving that 3D topology had an essential role in property prediction. Second, \(k\)-NAGCN with the new convolution mechanism achieved lower MAE than 3DGCN, proving that the proposed convolution mechanism could better model realistic atomic interactions. All three approaches achieved smaller MAE on \(\log_{10}(H_{cc} + 1)\) than \(\log_{10}(H_{cc})\). The result was predictable because \(\log_{10}(H_{cc} + 1)\) had smaller variation compared with \(\log_{10}(H_{cc})\). Considering both terms are one-to-one mapping, the \(\log_{10}(H_{cc} + 1)\) is a better choice for \(H_{cc}\) prediction task.

An ablation study was designed to evaluate the impacts of the new convolution mechanism and the 3D spatial information. The ablation study gradually removed the coordination transform, spatial information, and nearest neighbor convolution (the original CGCNN could be seen as the \(k\)-NAGCN removing all three components) from the \(k\)-NAGCN and compared the results with 3DGCN. The ablation study was tested on the ICSD dataset, including predicting large Li-site, percolation radius, and minimum bond length. Similarly, MAE, MAPE, and \(R^2\) scores were used to evaluate the performance. The results were summarized in Table 2.

```
| Method       | \(\log_{10}(H_{cc} + 1)\) | \(\log_{10}(H_{cc})\) |
|--------------|--------------------------|--------------------------|
| CGCNN [13]   | 0.117                    | 0.946                    |
| 3DGCN [26]   | 0.115                    | 0.950                    |
| \(k\)-NAGCN  | **0.101**                | **0.865**                |
```

TABLE 1: Performance of Henry’s constant prediction on CoRE2019 dataset. The performance is evaluated by Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and \(R^2\) score.
TABLE 2: Ablation study, regarding the impact of different k-NAGCN components on the ICSD dataset. The performance was evaluated in Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and $R^2$ score.

| Large Li-Site | Percolation Radius | Minimum Bond Length |
|--------------|--------------------|---------------------|
| **k-NAGCN**  | **MAE** | **MAPE** | **$R^2$** | **MAE** | **MAPE** | **$R^2$** | **MAE** | **MAPE** | **$R^2$** |
| without coordination transform $^*$ | 0.405 | 0.145 | 0.578 | 0.065 | 0.090 | 0.332 | 0.663 | 0.182 | 0.422 |
| without spatial information $^*$ | 0.360 | 0.118 | 0.665 | 0.060 | 0.079 | 0.403 | 0.688 | 0.188 | 0.477 |
| CGCNN [13] | 0.390 | 0.132 | 0.591 | 0.063 | 0.092 | 0.330 | 0.697 | 0.195 | 0.463 |
| 3DGCN [26] | 0.375 | 0.127 | 0.640 | 0.060 | 0.079 | 0.410 | 0.700 | 0.201 | 0.470 |

* The ablation steps are progressively applied, i.e., the model without spatial transform was afterward evaluated without coordination transform.

According to [33], the enlarged Li-site originates from the large local space within the crystal structural framework, and the percolation radius is defined as the maximum radius of a sphere that can percolate across at least one direction of the structure. Both definitions involve three-dimensional topology information. The results of the ablation study proved that the k-NAGCN algorithm benefited from the involvement of spatial information and the models without spatial information. The 3DGCN model also presented better performance than the CGCNN model with no spatial information.

Then we compared the k-NAGCN without spatial information with the CGCNN model. The proposed convolution mechanism, which modeled all neighbor atoms at one time, better fitted the physics laws and thus, outperformed the CGCNN model that only modeled pairwise atomic interactions. The k-NAGCN model without spatial information also outperformed the 3DGCN that contains spatial information by a small margin.

It was noticeable that the k-NAGCN without coordination transformation got the worst results among all competitors. This result revealed that the spatial information helped the performance only when the order of atoms was invariant to rotation and translation.

For a visual presentation of the results, figure 2 gives the scatter plot of the prediction results, where figure 2a presents the predictions of $H_{cc}$ on Core2019 dataset and figure 2b, figure 2c, and figure 2d presents the predictions of minimum bond length, percolation radius, and large Li-site on ICSD dataset, respectively.

The k-NAGCN model also shows good convergence and can avoid the overfitting problem well (see Figure 3).

The number of convolution layers is an important parameter that impacts the performance of the algorithm. However, the experimental results revealed that the number of convolution layers has a negligible effect on the results (see Figure 4). We also discovered that even the most simplified structure (two convolutional layers and one fully connected layer) fit the training data properly. Therefore, we used three convolutional layers and one fully connected layer throughout the rest of the experiments because the convolutional layers’ increment led to overfitting the training data.

One crucial hyper-parameter was the number of neighbor atoms $k$. While calculating the convolution (see equation 1), $k$ was vital because it determined the scale of the local structure, similar to the convolutional kernel size in a CNN. To evaluate the impact of the $k$, we repeated the experiments using the same setup with different scale parameters, which are 4, 8, 12, 16, and 32. The best average MAE achieved using different $k$ was summarized in Figure 5.

The number of neighbor atoms is another hyper parameter that could significantly impact the performance. The MAE is significantly larger when only a few neighbor atoms, e.g., four, were involved in the convolution. When $k$ was too small, the convolution did not consider the target atom’s whole local environment, and the model underperformed. The MAE was reduced while increasing the number and putting more atoms into the calculation. However, when we increased the size of the neighborhood to a specific range (e.g., 12 and 16), a further increment did not improve the performance. The performance decreased significantly when we increased the $k$ to 32. There were two reasons for the deterioration. First, the convergence was affected by the increased parameter number. Secondly, the atomic features were over-smoothed when the scale was too large.

One concern about the proposed method was the computational cost. The computational cost of the algorithm could be divided into two parts, graph building and convolution operation. The graph building included the data input, distance calculation, and coordination transformation. The coordination transformation was time-consuming and increased the total computational cost of k-NAGCN. However, the graph building was one-time computation and didn’t impact the calculation time. We counted the running time of the CGCNN and k-NAGCN on our platform (see Table 3). It could be seen that the k-NAGCN algorithm consumed more time on the graph construction process for both Core2019 and ICSD datasets.

We calculated the computational complexity of the graph convolutional layers of k-NAGCN and CGCNN separately. The batch size was set as 32, and the atom and edge feature-length were set as 64 and 41, respectively. The average number of atoms contained in the structure was set as 50, and the number of nearest neighbors was 12. The calculation ignores the activation functions and normalization layers. When both...
models contained 3 convolution layers, the computational complexity of the k-NAGCN is 1.572 GFLOPs, and the computational complexity of the CGCNN is 2.490 GFLOPS. The stastics in 3 reflected the above differences. The forward inference of the CGCNN model took a long time. The k-NAGCN had an advantage in computational efficiency during model training and cost more time for testing.

V. DISCUSSION

Accurate and rapid prediction of the physical properties of materials is a challenging problem. The recent significant advance in deep learning and the emergence of large-scale materials databases empower the development of machine learning technologies in the field. However, the variant material structure still stands in the way, hindering the immense
success of these technologies. The graph-based representation is concise and helpful in describing variant structures and effectively handling most materials analysis tasks. The undirected graph could hardly carry the higher-dimensional topology, which is vital to understanding the atoms’ interaction. This paper develops the k-NAGCN model that involves three-dimensional topological information.

The three-dimensional information makes the process more complex than the operation on the graph. One of the core problems is that the operation should be rotation invariant in the three-dimensional space. Inspired by the classic SIFT algorithm, we propose to rotate a new description that is rotation invariant. Moreover, Both [26] and [20] reported reduced calculation efficiency after involving the 3D information into consideration. The additional computational cost of the new convolution depends on the total number of atoms in the structure and the neighborhood scales \( s \), which is a constant. Thus, the computational complexity for description calculation is \( O(s \times n) \). The above computational cost does not affect the forward/backward process and has little impact on the training process.

The concept of scale is one significant difference between the existing model and the new model. Most existing methods loop over all the chemically bonded atom pairs and sum up the atomic interaction. The k-NAGCN eliminates the connected matrix and considers the proper scale parameter to prospect the corresponding local structure. One may question whether the fixed scale parameter would limit the model’s flexibility. However, the atomic interaction can be learned from data by considering the atom features and 3D atomic configuration, and it is unnecessary to be predefined.

VI. CONCLUSION AND FUTURE WORK

This work investigated three-dimensional geometric information usage in the graph-based model and proposed a novel k-NAGCN model. By comparing the experimental results, we found that 3D geometry played a vital role in modeling the effects of interatomic interactions on the physical properties of the materials. The proposed network structure better reflects real-world interaction than existing graph-based models that mainly ignore 3D topology.

Meanwhile, we developed an algorithm that generated rotation invariant descriptors for the distribution of local atoms. The invariant descriptor enhances the generalization of the prediction model by eliminating the ambiguity caused by rotation. The k-NAGCN model outperformed the existing graph-based methods in a group of challenging materials properties prediction tasks.

Our experiments proved that modeling materials structure as a 3D point cloud is feasible. The proposed model could be extended to general tasks for predicting the physical properties of materials and could be easily embedded in other models such as graph auto-encoders and graph generative networks. Meanwhile, introducing the scale parameter helps discover the most significant local environment for the atoms and may lead to novel meaningful chemistry structures.

The future work involves two parts. On the one hand, the proposed model will be used and validated on a series of property prediction tasks. It will also be used as the discriminator in the Generative Adversarial Network (GAN) that generates new MOFs. On the other hand, algorithms will be developed to enhance interpretability, which prompts the discovery of physic laws.

Table 3: Average time consumed for processing samples (ms).

|                | Graph Building | Training |
|----------------|----------------|----------|
|                | k-NAGCN       | CGCNN    | k-NAGCN | CGCNN |
| Core2019 [13] | 493.15        | 335.89   | 1.68    | 3.99  |
| ICSD [26]     | 120.7         | 113.06   | 1.31    | 1.54  |

REFERENCES

[1] A. Šimůnek and J. Vackář, “Hardness of covalent and ionic crystals: first-principle calculations,” Physical Review Letters, vol. 96, no. 8, p. 085501, 2006.
et al. [23] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *Journal of Physical: Condensed Matter*, vol. 14, no. 11, p. 2745, 2002.

I. Obot, D. Macdonald, and Z. Gasem, “Density functional theory (dft) as a powerful tool for designing new organic corrosion inhibitors. part 1: an overview,” *Corrosion Science*, vol. 99, pp. 1–30, 2015.

G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, “Towards an exact description of electronic wavefunctions in real solids,” *Nature*, vol. 493, no. 7432, pp. 365–370, 2013.

C. Gao, X. Min, M. Fang, T. Tao, X. Zheng, Y. Liu, X. Wu, and Z. Huang, “Innovative materials science via machine learning,” *Advanced Functional Materials*, vol. 32, no. 1, p. 2108044, 2022.

G. R. Schleider, A. C. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, “From dft to machine learning: recent approaches to materials science—a review,” *Journal of Physics: Materials*, vol. 2, no. 3, p. 035001, 2019.

J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, and M. Lei, “Machine learning in materials science,” *Infomat*, vol. 1, no. 3, pp. 338–358, 2019.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, “Benchmarking graph neural networks for materials chemistry,” *npj Computational Materials*, vol. 5, no. 1, p. 2101, 2019.

O. Allam, B. W. Cho, K. C. Kim, and S. S. Jang, “Application of dft-based machine learning for developing molecular electrode materials in li-ion batteries,” *RSC advances*, vol. 8, no. 69, pp. 39414–39420, 2018.

Y. Liu, T. Zhao, W. Ju, and S. Shi, “Materials discovery and design using machine learning,” *Journal of Materials*, vol. 3, no. 3, pp. 159–177, 2017.

K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “SchNet—a deep learning architecture for molecules and materials,” *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241722, 2018.

T. Xie and J. C. Grossman, “Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties,” *Physical review letters*, vol. 120, no. 14, p. 145301, 2018.

S. J. Greydanus, M. Dzumba, and J. Yosinski, “Hamiltonian neural networks,” 2019.

M. Hellenbrandt, “The inorganic crystal structure database (icsd)—present and future,” *Crystallography Reviews*, vol. 10, no. 1, pp. 17–22, 2004.

A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder et al., “Commentary: The materials project: A materials genome approach to accelerating materials innovation,” *APL materials*, vol. 1, no. 1, p. 011002, 2013.

S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühi, and C. Wolverton, “The open quantum materials database (oqmd): assessing the accuracy of dft formation energies,” *npj Computational Materials*, vol. 1, no. 1, pp. 1–8, 2015.

T. M. Mitchell, “Machine learning and data mining,” *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

E. D. Cubuk, A. D. Sendek, and E. J. Reed, “Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data,” *The Journal of chemical physics*, vol. 150, no. 21, p. 214701, 2019.

J. Hoffmann, L. Maestri, Y. Sawada, J. Tang, J. M. Sellier, and Y. Bengio, “Data-driven approach to encoding and decoding 3-d crystal structures,” *arXiv preprint arXiv:1909.00949*, 2019.

X. Zheng, P. Zheng, and R.-Z. Zhang, “Machine learning material properties from the periodic table using convolutional neural networks,” *Chemical science*, vol. 9, no. 44, pp. 8426–8432, 2018.

S. Sanyal, J. Balachandran, N. Yadati, A. Kumar, P. Rajagopalan, S. Sanyal, and P. Talukdar, “M-gcnn: Integrating crystal graph convolutional neural network with multitask learning for material property prediction,” *arXiv preprint arXiv:1811.05660*, 2018.

Y. Zhang and Q. Yang, “A survey on multi-task learning,” *arXiv preprint arXiv:1707.08114*, 2017.

C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, “Graph networks as a universal machine learning framework for molecules and crystals,” *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

J. Lee and R. Asahi, “Transfer learning for materials informatics using crystal graph convolutional neural network,” *Computational Materials Science*, vol. 190, p. 110314, 2021.

H. Cho and I. S. Choi, “Three-dimensionally embedded graph convolutional network (3dgcnn) for molecule interpretation,” *arXiv preprint arXiv:1811.09794*, 2018.

Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, “Learning atoms for materials discovery,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, pp. E6411–E6417, 2018.

D. Harris and S. Harris, *Digital design and computer architecture*. Morgan Kaufmann, 2010.

D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, Conference Proceedings, pp. 1150–1157.

S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, “Python materials genomics (pymatgen): A robust, open-source python library for materials analysis,” *Computational Materials Science*, vol. 68, pp. 314–319, 2013.

D. Koks, *Explorations in mathematical physics: the concepts behind an elegant language*. Springer Science & Business Media, 2006.

J. L. Soto, P. W. Fisher, A. J. Glessner, and A. L. Myers, “Sorption of gases in molecular sieves. theory for henry’s constant,” *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, vol. 77, no. 1, pp. 157–168, 1981.

X. He, Q. Bai, Y. Liu, A. M. Nolan, C. Ling, and Y. Mo, “Crystal structural framework of lithium super-ionic conductors,” *Advanced Energy Materials*, vol. 9, no. 43, p. 1902078, 2019.

N. J. English and D. G. Carroll, “Prediction of henry’s law constants by a quantitative structure property relationship and neural networks,” *Journal of chemical information and computer sciences*, vol. 41, no. 5, pp. 1150–1161, 2001.

BOYU ZHANG Boyu Zhang is a Postdoctoral Fellow in the Department of Mathematics and Statistics and Institution for Modeling Collaboration and Innovation, University of Idaho. He holds a Ph.D. in Computer Science from the Harbin Institute of Technology. His research interests include Machine Learning, Computer Vision, and Medical Image Analysis.

MUSHEN ZHOU Musen Zhou received his Bachelor’s degree in Chemical Engineering from Nanjing Tech University in 2017. He is currently a Ph.D. candidate at the University of California, Riverside, advised by Prof. Jianzhong Wu. His current research interests are using physics- and data-based modeling to inverse design nanoporous materials for gas separation and energy storage.

JIANZHONG WU Professor Jianzhong Wu is a faculty member of the Department of Chemical and Environmental Engineering and a cooperating faculty of the Mathematics Department at the University of California, Riverside. He received his Ph.D. from the University of California, Berkeley, and MS and BS degrees from Tsinghua University, Beijing. Dr. Wu’s research has been directed at the development and application of statistical-mechanical methods for predicting the microscopic structure and physicochemical properties of inhomogeneous fluids, soft matter, and biological systems.
FUCHANG GAO  Fuchang Gao is a Professor of Mathematics at the University of Idaho, with a Ph.D. degree from the University of Connecticut. Dr. Gao works on the Theory of Approximation, Theory of Probability, and statistical and data science applications.
APPENDIX A THE CALCULATION OF HENRY’S CONSTANT

Here, we present the Henry’s constant in the dimensionless format calculated as followings [32],

$$H^*_{cc} = \frac{1}{V} \int e^{-\beta V^{ext}(r)} dr$$

(A1)

where $\beta = 1/(k_B T)$, $R$ represents the gas constant, $k_B$ stands for the Boltzmann constant. $T$ is the absolute temperature, and $V^{ext}$ is the external potential. In this work, we consider hydrogen adsorption in MOF structures where the external potential is mainly contributed by van der Waals interactions. The Lennard-Jones 12-6 potential $V_{LJ}$ is therefore used to describe the pairwise additive potential for the interaction of a hydrogen molecule with each atom inside the MOF structure with the cutoff distance of 12.9 Å.

$$V_{LJ} = 4\varepsilon \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6$$

(A2)

where $r$ is the pairwise distance between gas molecules and atoms in the MOFs, and $\sigma$ and $\varepsilon$ are the LJ size and energy parameters. In this work, universal force field (UFF) is used for all atoms in MOFs. For hydrogen molecule, $\sigma_{H_2} = 0.296$ nm and $\varepsilon_{H_2}/k_B = 34.2$ K. For interactions between different kinds of atoms, the Lorentz-Berthelot mixing rules were used.

APPENDIX B THE REMOVAL OF MOF STRUCTURES WITH EXTREME VALUES

The original CoRE2019 MOF database contains 12763 MOF structures. According to our calculation of Henry’s constant, the range of Henry’s constant covers more than 30 orders of magnitude (as shown in Figure B1). After examining the MOF structure files, we found the broad range of Henry’s constants was due to some unreasonable MOF structures. In Figure B1, we can see that most MOFs distribute around unit Henry’s constant after database cleanup. There are a small portion of MOFs with exceptionally high Henry’s constants in Figure B1.

A closer examination reveals that atoms are unreasonably close to each other in these structures. For example, in MOF (RUBLEH) with the highest Henry’s constant, the distance of oxygen-oxygen there is only 0.366 Å, which is even shorter than the shortest covalent bond (hydrogen-hydrogen 0.74 Å). The overlapped atom position would lead to unreasonably high attraction and unreasonably high Henry’s constant.

After getting rid of MOF structure with potentially overlapped atoms (with the smallest atom distance smaller than 0.8 Å), the distribution of Henry’s constant is more reasonable, as shown in Figure B2. Most MOFs with exceptionally high Henry’s constant are filtered out.

FIGURE B2: Distribution of Henry’s constant for MOFs without potentially overlapping atoms in CoRE2019 MOF database.

It is also worth noticing that some Henry’s constant values are also extremely low (on the order of $10^{-10}$). The low Henry’s constant for those MOFs is due to the relatively small pore size or tightly confined geometry (XEKUO). Compared to other MOFs with similar void fractions, XEHUO has a relatively thin structure. Therefore, most interactions inside the MOF are repulsive, resulting in an ultra-low Henry’s constant. For example, compared with XEJJOM (void volume of 253 Å³), XEKUO (void volume of 185 Å³) has 0 surface area while XEJJOM has a surface of 9.40 Å² when using test particle with a diameter of 3.68 Å(size of nitrogen molecule). After removing MOFs with surface areas less than 50 Å² (meaning that there will be very little space and very unlikely for surface adsorption to happen), we filtered out all the MOFs with extremely low Henry’s constant. The cleaned data set contains 10136 MOF structures. The values of Henry’s constant still cover the several order of magnitudes. However, we believe they may be reachable. The distribution of Henry’s constants of the cleaned dataset is shown in Figure B3.
FIGURE B1: Distribution of Henry’s constants for hydrogen gas in 12763 MOFs in original CoRE2019 MOF database (left) and 10136 MOFs after cleanup (right).

FIGURE B3: Distribution of Henry’s constants for hydrogen gas in MOFs without structures containing potentially overlapped atoms and extremely small surface area (left). Xe/Kr selectivity for Henry’s constant from the literature (right).