Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective

Chengqiang Lu,[†] Qi Liu,[†*] Chao Wang,[†] Zhenya Huang,[†] Peize Lin,[‡] Lixin He[‡]
[†]Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China
[‡]Key Laboratory of Quantum Information, University of S&T of China
{qiliuql, helx}@ustc.edu.cn, {lunar, wdyx2012, huangzhy, linpz}@mail.ustc.edu.cn

Abstract

Predicting molecular properties (e.g., atomization energy) is an essential issue in quantum chemistry, which could speed up much research progress, such as drug designing and substance discovery. Traditional studies based on density functional theory (DFT) in physics are proved to be time-consuming for predicting large number of molecules. Recently, the machine learning methods, which consider much rule-based information, have also shown potentials for this issue. However, the complex inherent quantum interactions of molecules are still largely underexplored by existing solutions. In this paper, we propose a generalizable and transferable Multilevel Graph Convolutional neural Network (MGCN) for molecular property prediction. Specifically, we represent each molecule as a graph to preserve its internal structure. Moreover, the well-designed hierarchical graph neural network directly extracts features from the conformation and spatial information followed by the multilevel interactions. As a consequence, the multilevel overall representations can be utilized to make the prediction. Extensive experiments on both datasets of equilibrium and off-equilibrium molecules demonstrate the effectiveness of our model. Furthermore, the detailed results also prove that MGCN is generalizable and transferable for the prediction.

Introduction

Predicting molecular properties, such as atomization energy, is one of the fundamental issues in quantum chemical science. Indeed, it has attracted much attention in relevant fields of physics, chemistry and computer science, since it speeds up the societal and technological progress in the application of discovering substances with desired characteristics, such as drug design with specific target and new material manufacture (Becke 2007; Oglic, Garnett, and Gärtner 2017).

In the literature, density functional theory (DFT) plays an important role in physics for molecular property prediction. It holds a common statement that the quantum interactions between particles (e.g., atom) create the correlation and entanglement of molecules which are closely related to their inherent properties (Thouless 2014). Along this line, many quantum mechanical methods based on DFT have been developed to model the quantum interactions of molecules for the prediction (Hohenberg and Kohn 1964; Kohn and Sham 1965). However, DFTs are computationally costly since they usually use specific functions to determine the interactions of particles, which proves to be extraordinarily time consuming. For example, experimental results indicated that it took nearly an hour to predict the properties of merely one molecule with 20 atoms (Gilmer et al. 2017). Obviously, it is unacceptable to make prediction on large number of molecules in chemical compound space. Therefore, it is necessary to find more effective solutions.

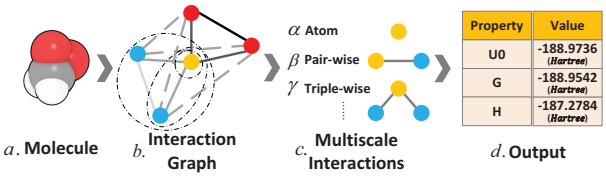

Figure 1: Illustration of the process of a molecule (CH_2O_2) via our method.

Recently, inspired by the remarkable success of machine learning in many tasks including computer vision, natural language processing, natural and social science (Karpathy et al. 2014; He et al. 2016; Huang et al. 2017; Zhu et al. 2018; Liu et al. 2018), researchers have shown the potentials of these data-driven techniques for molecular property prediction (Faber et al. 2017; Schütt et al. 2017a). Generally, these studies mainly rely on rule-based feature engineering (e.g., bag of atom bonds) or treat molecules as grid-like structures (e.g., 2D images or text). However, few of them directly take the inherent quantum interactions of molecules into consideration, causing severe information loss, which makes the molecular property prediction problem pretty much open.

Unfortunately, there are many technical and domain challenges along this line. First, there are highly complex quantum interactions, such as distracted attraction, exchange repulsion and electrostatic interaction in molecules, especially in the large molecules (Kollman 1985). It is hard to model them with analytical methods. Second, compared with traditional tasks including computer vision, the amount of labeled molecule data is significantly limited, which requires a generalizable approach for the prediction. Last but not least, in practice, we are often provided with labeled data of small

*Contact author.

and medium molecules except large molecules since the calculation of them are expensive. Thus, it is necessary to notice this unbalancedness to propose a transferable solution for property prediction of large molecules using the model trained on smaller ones.

To address these challenges, in this paper, we propose a well-designed Multilevel Graph Convolutional Neural Network (MGCN) for predicting properties of molecules by directly incorporating their quantum interactions. Figure 1 demonstrates the process of our approach. Specifically, we first represent each molecule as an interaction graph, which could preserve its internal structure without information loss. Then we propose a hierarchical graph convolutional neural network to model the multilevel quantum interactions based on the graph-like molecular structures. Here, we follow the DFT theory that the quantum interactions could be transformed at different levels, i.e., atom-wise refers to the inherent influence of each atom (e.g., oxygen), atom-pair refers to the interaction between two atoms, atom-triple means the correlation among three atoms, and so on. Thus, our proposed graph network incorporates hierarchical layers of point-wise, pair-wise, triple-wise, etc to extract representations of the multilevel interactions, respectively. Finally, the overall interaction representation from all levels could be utilized to make the property (e.g., atomization energy) prediction. We conduct extensive experiments on both datasets of equilibrium and off-equilibrium molecules, where the experimental results shows the effectiveness of our proposed approach. Moreover, as MGCN could naturally pass the interaction information of molecules level by level, which also proves the superior ability of generalizability and transferability.

Related Work

Generally, the related work of our research could be classified into the following three categories.

Density Functional Theory. Molecular property prediction problem has been studied for a long time in physics, chemistry and material science (Wang and Hou 2011). In the literature, density functional theory (DFT) is the most popular method, which plays a vital role in making the prediction, and could date back to 1960s (Hohenberg and Kohn 1964; Kohn and Sham 1965; Lawless and Chandrasekara 2002). Generally, it states that the quantum interactions between particles (e.g., atoms) create the correlation and entanglement of molecules which are closely related to their inherent properties (Thouless 2014). Following this theory, many DFT based methods, such as B3LYP, were proposed, which mapped the quantum interactions of molecules onto every single particles, for predicting the properties (Yanai, Tew, and Handy 2004). However, the complexity of DFT could be approximated as $\mathcal{O}(N^3)$, where N denotes the number of particles. Therefore, it is time-consuming in the experiments and unacceptable for the prediction when facing large number of molecules (Gilmer et al. 2017).

Traditional Machine Learning Methods. To find more efficient solutions for molecular property prediction, researchers have attempted to leverage various machine learning models, such as kernel ridge regression, random forest and Elastic Net (Faber et al. 2017; Zou and Hastie 2005; McDonagh et al. 2017). Generally, they rely on rule-based hand crafted features using the domain knowledge of physics and chemistry, including bag of bonds, coulomb matrix, and histogram of distances, angles and dihedral angles (Huang and von Lilienfeld 2016; Hansen et al. 2015; Montavon et al. 2012). Although some superior experimental results have been achieved, these traditional machine learning methods take manual feature engineering, which requires much domain expertise. Thus, they are often restricted in practice.

Deep Neural Networks. Compared to traditional machine learning models, deep neural networks hold a superiority of automatic feature learning, which have achieved great success in many applications, such as speech recognition (Zhu et al. 2016), computer vision (LeCun, Bengio, and others 1995) and natural language processing (Collobert and Weston 2008). With this ability, researchers have noticed the potentials of these deep methods for molecular property prediction. Along this line, convolutional neural network based models were proposed, where they represented each molecule as grid-like structures, such as image (Goh et al. 2017), string (Gómez-Bombarelli et al. 2018), and sphere (Boomsma and Frellsen 2017). For example, Goh et al. (2017) converted molecular diagrams into 2D RGB images and proposed the ChemNet for the prediction. However, this grid-like transformation usually caused information loss of the molecules which lied in non-Euclidean space, where the internal spatial and distance information of atoms were not fully considered (Bronstein et al. 2017). Therefore, some works operated the molecule as a atom graph and developed graph convolutional neural networks for the property prediction (Schütt et al. 2017b; Gilmer et al. 2017). For instance, Schutt et al. (2017b) proposed the deep tensor neural network that captured the representation of each atom node in molecules. Shang et al. (2018) further introduced attention mechanism for characterizing the edge information to improve the prediction.

Our work improves the previous studies as follows. First, we propose the multilevel graph network to directly model the multilevel quantum interactions of molecules from hierarchical perspectives (i.e., atom-wise, pair-wise, triple-wise, etc), which developed the graph modeling for molecular property prediction. Second, our work could pass the interaction information level-by-level, which benefits more practical scenarios, i.e., generalizability of limited data and transferability of unbalanced data.

Multilevel Graph Convolutional Network

In this section, we first formally introduce the molecular property prediction problem. Then we describe our Multilevel Graph Convolutional Network in detail.

Problem Statement

Given a molecule, it is natural to represent it as graphs without the loss of information, where vertices represent atoms and edges represent chemical bonds. Thus, a molecule is denoted by $\mathcal{G}\langle\mathcal{V},\mathcal{E}\rangle$, and in the setting of molecular structure,

$\mathcal{V}$ is a set of atoms with $|\mathcal{V}| = N$. We regard the graph as a complete undirected graph following the assumption that every atom has the interactions with others so that the set of edges satisfies that $|\mathcal{E}| = N(N-1)/2$. Here each $\mathcal{E}$ contains two kinds of information, namely edge type and spatial information, respectively. Our target is to construct a regressor to predict the properties of molecules. Formally, we can define the problem as:

$$g(f(\mathcal{G})) = y, \tag{1}$$

where y is the target property to predict and the middle function $f : \mathcal{G} \rightarrow \mathbb{R}^{N\times D}$ is used to learn representations of atoms. Then g converts the obtained features to final result.

The multilevel interactions widely exist in the graph structures. In the field of molecule, physical experts design the different symmetry functions to describe the atomic environment by considering the interactions at varied levels (Behler 2014). Inspired by this idea, we model quantum interactions in molecules by representing the interactions between two, three, and more atoms level by level to demystify the complexity of molecular interactions. In the next subsection, we will introduce our Multilevel Graph Convolutional Network (MGCN) in detail.

Network Architecture

Overview. The entire architecture could be split into three parts in a high-level discussion except for thse input. The initial input is a graph which consists of a list of atoms and a Euclidean distance matrix of the molecules. The preprocessing part includes embedding layer and Radial Basis Function (RBF) layer. The embedding layer generates atom and edge embeddings while Radial Basis Function (RBF) layer converts the distance matrix to a distance tensor. The next part of MGCN are several interaction layers that aim to learn different node representations in different levels. The last phase is the readout layer that outputs the final result.

Embedding Layer. Atoms and bonds are the basic elements in a molecule. Thus, to model interactions with as less information loss as possible, we present an embedding layer to directly embed vertices and edges of a graph into vectors. Each atom in a molecule is represented as a vector $\boldsymbol{a}^0 \in \mathbb{R}^D$ initially. Therefore, the vertices in the entire molecular are denoted as a matrix $A^0 \in \mathbb{R}^{N\times D}$ and $\boldsymbol{a}_i^0$ indicates the atom embedding of i-th atom in a molecule. The atoms that have the same number of protons in their atomic nuclei share the same initial representation which is called the atom embedding here. Taking CH_2O_2 as an example, there are five atoms and different kinds of atoms are labeled with different colors in the input part of Figure 2. After the process of embedding layer, we get a $5\times D$ matrix and the rows that are related to the atoms of the same type share the same value. The atom embeddings of all chemical elements are generated randomly before training. The initialization of pair-wise embeddings $\boldsymbol{e} \in \mathbb{R}$ is similar to atom embeddings (see the preprocess part of Figure 2). Thus we get $E \in \mathbb{R}^{N\times N\times D}$, and the edges connecting the same set of atoms have the same initial edge embedding. Specifically, $\boldsymbol{e}_{ij}$ indicates the edge embedding of the bond between i-th atom and j-th atom. The representations generated by the embedding layer are only related to the inherent property of isolated atoms and bonds. The interaction terms are modeled in later subnetwork.

Radial Basis Function Layer. The spatial information influences the degree of interactions between nodes and we use the RBF Layer to convert these information to robust distance tensors for further utilization. First of all, we reform the raw coordinates of atoms to distance matrix to remove the disturbance of selection of coordinate frame. Secondly, Radial Basis Functions are applied to convert the distance matrix to a distance tensor.

RBF is a widespread kernel method which originally was invented to generate function interpolation (Broomhead and Lowe 1988). Its variant was proved to be advantageous to create fingerprint-like descriptor of molecules (Li, Han, and Wu 2018). Here we use RBF to spread the 2D inter-atomic distance matrix to a 3D representation. Given a set of K central points $\{\mu_1, \ldots, \mu_K\}$, the single data point x, namely one pair-wise distance in the molecule, will be processed as:

$$RBF(x) = \overset{K}{\underset{i=1}{\frown}} \mathrm{h}(\|x - \mu_i\|). \tag{2}$$

Here the notation $\frown$ means concatenation, and we take Euclidean distance as the norm. As for radial basis function h, we take Gaussian $\exp(-\beta\|x-\mu_i\|^2)$ following the suggestion in (Schütt et al. 2017a) to avoid the long plateau on the initial phases of the training procedure. k central points are picked evenly in the range from the shortest to the longest edge among the entire dataset. Therefore all distances in the dataset will be covered.

Through the non-linear transformation, the representations of distances between nodes become more robust. Furthermore, more additional interpretation is introduced by radial basis function layer than simple multi-layer perceptron. After the RBF layer, we create the pair-wise distance tensors $D \in \mathbb{R}^{N\times N\times K}$, and d_{ij} denotes the distance tensor between i-th atom and j-th atom.

Interaction Layer. To model the multilevel molecular structure with all the conformation and spatial information embedded through previous layers, we construct the interaction layer which is a crucial component of our model. Considering that the quantum interactions in molecules could be transformed at different levels (i.e., atom-wise, atom-pair, atom-triple, etc), our interaction layer is designed by the hierarchical architecture level by level. Specifically, in the l-th interaction layer, we define the edge representation e_{ij}^{l+1} and atom representation a_i^{l+1} as:

$$\boldsymbol{e}_{ij}^{l+1} = \mathrm{h}_e(\boldsymbol{a}_i^l, \boldsymbol{a}_j^l, \boldsymbol{e}_{ij}^l), \tag{3}$$

$$\boldsymbol{a}_i^{l+1} = \sum_{j=1, j\neq i}^{N} \mathrm{h}_v(\boldsymbol{a}_j^l, \boldsymbol{e}_{ij}^l, \boldsymbol{d}_{ij}), \tag{4}$$

where h_e is used to update edge representation and h_v is the function that collects the message from the neighbours of the i-th atom to generate $\boldsymbol{a}_i^{l+1}$.

With this hierarchical modeling, MGCN could effectively preserve the structure of each molecules and describe its quantum interactions. Specifically, in the first layer, $\boldsymbol{a}_i^0$ denotes the atom embedding that show the inherent properties of certain chemical elements. As the forward inference

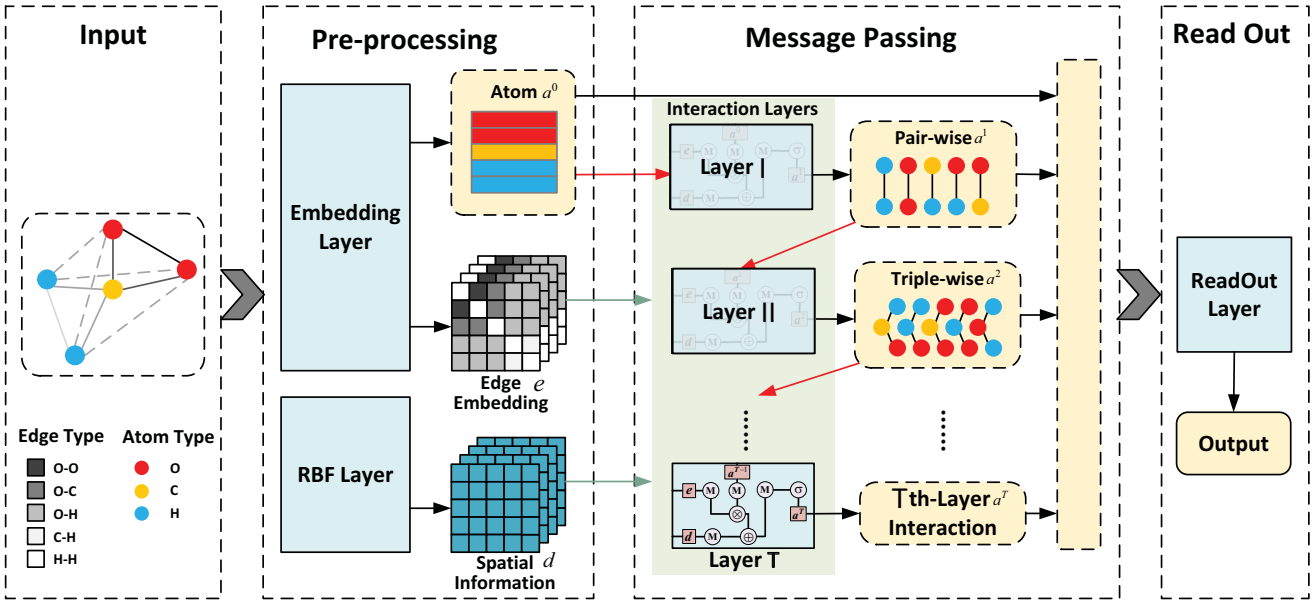

Figure 2: The architecture of the entire MGCN.

steps, $\boldsymbol{a}_i^1$ involves the first-order neighbour node and spatial information with the message passed by $\boldsymbol{a}^0$, e and d. In a similar way, $\boldsymbol{a}_i^2$ represents the triple-wise interactions, $\boldsymbol{a}_i^3$ indicates the interactions between four nodes and so on. As shown in Figure.2, after each interaction layer, we obtain the representations of atoms that reflect the higher-order interactions thanks to decomposition of molecule.

The update function h_e is calculated as:

$$\mathrm{h}_e = \eta \boldsymbol{e}_{ij}^l \oplus (1-\eta) W^{ue} \boldsymbol{a}_i^l \odot \boldsymbol{a}_j^l. \tag{5}$$

where η is the hyper-parameter that controls the influence of former pair-wise information (default value is 0.8). Here $\odot$ and $\oplus$ denote the element-wise dot and plus respectively. In this way, The edge embedding is corrected by the related atomic representations of former interaction layer.

The function h_v applies the message passing operation to create the atom representation at a higher order. The distance tensor d_{ij} here controls the magnitude of impact in each pair of atoms and the edge embedding e_{ij} provides the extra bond information. Thus, it combines the information of nodes, edges, and space, more formally:

$$\mathrm{h}_v = \sigma(W^{uv}(\mathrm{M}^{fa}(\boldsymbol{a}_j^l) \odot \mathrm{M}^{fd}(\boldsymbol{d}_{ij}) \oplus \mathrm{M}^{fe}(\boldsymbol{e}_{ij})), \tag{6}$$

where σ is a tanh activation function, W^{uv} is a weight matrix. The notation $\mathrm{M}(\boldsymbol{x})$ refers to a dense layer that $\mathrm{M}(x) = W\boldsymbol{x} + \boldsymbol{b}$ with the input $\boldsymbol{x}$ for simplicity. M^{fa},M^{fd},M^{fe} are dense layers here.

Readout Layer. After the interaction layers, we get atom representations at different levels. In the last phase, we construct a readout layer to make the final prediction utilizing these features more clearly.

First of all, we aggregate the various atom representations to obtain the final vertex feature map as following:

$$\boldsymbol{a}_i = \overset{T}{\underset{k=0}{\frown}} \boldsymbol{a}_i^k, \tag{7}$$

where T indicates the number of interaction layers and $\frown$ means concatenation.

Secondly, we need to predict the property of molecule with the multilevel representations of each atoms. Fortunately, the molecular properties satisfy additivity and locality. For example, to predict the energetic property, we can model potential energy surfaces as follows (Behler 2014; Cubuk et al. 2017):

$$E = \sum_i^N \sum_j^N E_{ij}, \tag{8}$$

where E is the total energy and E_{ij} indicates the part of energy related to the bond between i-th and j-th atom ($i \neq j$). Besides, E_{ii} could be regarded as the partial energy that mapped to i-th atom. Along this line, we can process the representations separately and then sum them up:

$$\hat{y} = \sum_{i=1}^N W^{r_2^a} \sigma(\mathrm{M}^{r_1^a}(\boldsymbol{a}_i)) + \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{j=1}^N W^{r_2^e} \sigma(\mathrm{M}^{r_1^e}(\boldsymbol{e}_{ij})), \tag{9}$$

where σ is the activation function, more specifically, the softplus function. The former term refers to the contribution of quantum interactions that mapped to each atom. Additionally, the latter term denotes the edge-related contribution that can not be mapped to single particle. Since the atom-related interactions account for vast majority of molecular interactions, the latter term is nuanced. Therefore, when the amount of data is small, we tend to ignore the latter term.

To train this model, we use the Root-Mean-Square Error (RMSE) as our loss function:

$$\ell(\hat{y}, y) = \sqrt{|\hat{y} - y|^2}, \tag{10}$$

where $\hat{y}$ denotes the predictive value and y is the true value.

Discussion on MGCN

Generalizability. In the field of chemistry, the set of all possible molecules in unexplored regions is called chemical space. One of the famous chemical space project (Ruddigkeit et al. 2012) collected 166.4 billion molecules while merely 134k samples of them were labeled (Ramakrishnan and von Lilienfeld 2015). Therefore, the generalization ability of enabling accurate prediction with the limited dataset is indeed essential in our task.

In the design of our model, we decide to use the distance tensors D as the form of spatial information instead of coordinates of atom. Accordingly, MGCN enforces rotation and translation invariance. Henceforth, the representation learned by MGCN is more general and would not be confused with the same molecule in different orientations.

Moreover, we perform element-wise operations in interaction layers (equation(5) and (6)) to generate representations and process the representations of each atom respectively. Under those circumstances, the prediction made by our model is irrelevant to sequence of atoms. The index invariance enhances the generalization ability of MGCN.

Additionally, we utilize some normalization techniques such as dropout to prevent overfitting, which also benefits generalizability of our model.

In brief, our model is generalizable which is particular important for molecular property prediction where the amount of training data is limited.

Transferability. Since the expensive computational cost is a critical bottleneck which limits capabilities to calculate the properties of large molecules, most open data are small and medium molecules and the amount of large molecules is small. Therefore, the ability of transferring the knowledge learned from small molecules to larges ones could help us deal with the data-hungry of big molecules.

The atom/edge embeddings generated by embedding layer are only in regard to the type of atoms and edges and irrelevant to the specific molecular structure and spatial information. As a result, the chemical-domain knowledge learned in the embeddings is universal in the molecular system no matter small or large molecules. Then, in our mulitilevel phase, we use the embeddings to generate the representation in deeper level, e.g., pair-wise and triple-wise. Although small and large molecules are different in the distribution of atoms and bonds, their interactions in different levels are similar. Consequently, with the general embeddings and similar interaction mechanism, MGCN could infer the higher-level representations to predict property and maintain a certain accuracy. Therefore, our model that trained on the small molecules could obtain competitive performance in the prediction of larger molecules.

Rather than applying the model trained on small molecules to big molecules directly, another way to transfer the knowledge is using pre-trained embeddings. To train a model in large molecules, we could initialize this model with the atom and edge embeddings of another model that was trained on small molecules. The pre-trained embeddings could speed up the convergence and improve the accuracy, because the domain knowledge in embeddings learned from small molecules is still meaning suitable to big molecules.

Along this line, this model is capable of transferring the knowledge of small molecules to large molecules and tackle the structural shortage of data.

Time Complexity. The time complexity of our MGCN model is $\mathcal{O}(N_a^2)$ since the calculation of h_e and h_v (equation (5) and (6)) are independent of molecular size. Here N_a indicates the number of atoms in a molecule.

Experiments

We conduct experiments to demonstrate the effectiveness of MGCN from various aspects: 1) predictive performance; 2) the effectiveness of multilevel structure; 3) the validation of generalizability; 4) the verification of transferability; 5) the influence of varied number of interaction layers.

Datasets

QM9. The QM9[1] dataset (Ramakrishnan et al. 2014) is perhaps the most well-known benchmark dataset which contains 134k equilibrium molecules with their 13 different properties. All of the relaxed geometries and properties for all the 134k molecules are calculated by DFT. The DFT error is the empirical inaccuracy estimation of DFT based approaches (Faber et al. 2017). The QM9 dataset also provides the chemical accuracy which is generally accepted by the chemistry community as a relatively ideal accuracy.

ANI-1. The ANI-1[2] dataset provides access to the total energies of 20 million off-equilibrium molecules which is 100 times larger than QM9.

Experimental Setup

We use mini-batch stochastic gradient descent (mini-batch SGD) with the Adam optimizer (Kingma and Ba 2014) to train our MGCN. The batch size is set to 64 and the initial learning rate is $1e^{-5}$. For all 13 properties of QM9, we pick 110k out of 130k molecules randomly as our training set that accounts for about 84.7% of the entire dataset. With the rest of the data, we choose half of them as the validation set and the other half as the testing set. As for the much larger ANI-1, we randomly choose 90% samples for training, 5% samples for validation and 5% for testing. We select Mean Absolute Error (MAE) as our evaluation metrics for the convenience of comparison with baselines (Faber et al. 2017).

Baselines

We compare our model with the 7 baseline methods that could be categorized into two groups.

The first group consist of 3 traditional ML models using hand-engineered features derived from the molecular literature (Faber et al. 2017; Huang and von Lilienfeld 2016; Hansen et al. 2015). These ML models include Random Forest (RF) and Kernel Ridge Regression (KRR). The hand-craft features include Bag of Bonds (BOB), Bond-Angle Machine Learning (BAML) and "Projected Histograms" (HDAD). We imply the combination of X regressor and Y representation with the notation X+Y. Thus, these

[1] http://www.quantum-machine.org/datasets/#qm9

[2] https://www.nature.com/articles/sdata2017193

Table 1: Predictive accuracy of different models in QM9

Properties Unit	U_0 eV	U eV	G eV	H eV	C_v cal/molK	$\varepsilon_{\text{HOMO}}$ eV	$\varepsilon_{\text{LUMO}}$ eV	$\Delta\varepsilon$ eV	ω_1 cm^{-1}	ZPVE eV	$\langle R^2 \rangle$ Bohr^2	μ Debye	α Bohr^3
DFT Error	0.1	0.1	0.1	0.1	0.34	-	-	-	28	0.0097	-	0.1	0.4
Chemical Acc.	0.043	0.043	0.043	0.043	0.05	0.043	0.043	0.043	10	0.00122	1.2	0.1	0.1
RF+BAML	0.2000	-	-	-	0.451	0.1070	0.1180	0.1410	2.71	0.01320	51.10	0.434	0.638
KRR+BOB	0.0667	-	-	-	0.092	0.0948	0.1220	0.1480	13.20	0.00364	0.98	0.423	0.298
KRR+HDAD	0.0251	-	-	-	0.044	0.0662	0.0842	0.1070	23.10	0.00191	1.62	0.334	0.175
GG	0.0421	-	-	-	0.084	0.0567	0.0628	0.0877	6.22	0.00431	6.30	0.247	0.161
enn-s2s	0.0194	0.0194	0.0168	0.0189	0.040	0.0426	0.0374	0.0688	1.90	0.00152	0.18	**0.030**	0.092
DTNN	0.0364	0.0377	0.0385	0.0357	0.089	0.0982	0.1053	0.1502	4.23	0.00312	0.30	0.257	0.131
SchNet	0.0134	0.0189	0.0196	0.0182	0.067	0.0507	**0.0372**	0.0795	3.83	0.00172	0.27	0.071	0.073
MGCN	**0.0129**	**0.0144**	**0.0146**	**0.0162**	**0.038**	**0.0421**	0.0574	**0.0642**	**1.67**	**0.00112**	**0.11**	0.056	**0.030**

Table 2: Predictive accuracy of different models in ANI-1

Methods	DTNN	SchNet	MGCN
MAE	0.113	0.108	**0.078**

three baselines are denoted by RF+BAML, KRR+BOB, KRR+HDAD. These models achieve the best performance in the prediction of one or more properties among all 30 combinations of regressors and features (Faber et al. 2017).

The second group contain 4 deep neural networks. They are gated graph network (GG, Kearnes et al. 2016), edge neural network with set-to-set (enn-s2s, Gilmer et al. 2017), deep tensor neural network (DTNN, Schütt et al. 2017b) and SchNet (Schütt et al. 2017a). These models are proved to be competitive in the molecular property prediction. Noting that DTNN and SchNet only provide their experimental results in the prediction of property U_0 for the molecules in QM9, thus we complete the rest of the experiments. Besides, all of other numerical results of the baselines are extracted from their works directly.

Experimental Results

Predictive performance. We compare our model with the baseline models mentioned above in two datasets. In Table 1, we provide the MAE of baselines and our approach as well as DFT error and chemical accuracy for all 13 properties. Table 2 shows the performance comparison in ANI-1.

As illustrated in Table 1, MGCN gets the best performance in 11 out of 13 properties, and 11 of them exceed the chemical accuracy. Our model is able to improve the performance upon state-of-the-art. Another observation is that the deep neural networks (GG, enn-s2s, DTNN, SchNet and MGCN) outperform the models that using hand-craft features comprehensively. In the experiment in ANI-1, we choose the state-of-the-art models (DTNN and SchNet) as comparison. As shown in Table 2, the accuracies in ANI-1 is lower than in QM9, and there are possible two reasons. First, the force in equilibrium molecules of QM9 is negligible, while in off-equilibrium molecules of ANI-1, this factor increases the complexity of quantum interactions. Second, the 100 times larger size of ANI-1 than QM9 makes it more difficult to fit. Even though, our model still achieves satisfactory accuracy and outperform other methods.

In brief, our model attains the best performance benefiting from the mulitilevel interaction modeling, and the results prove that our model could handle both equilibrium and off-equilibrium molecules and is capable to fit the large dataset. Considering that most previous work have no experiment in the ANI-1 dataset, we chose QM9 as our default dataset in the rest of our experiments.

Effectiveness of multilevel interactions. In the molecular system, with the increase of the number of atoms in a molecule, the complexity of quantum interactions will grow exponentially. In consequence, it is much harder to model the interactions of molecules if the size of them is larger.

Figure 3 shows the MAE of predictions as the function of the number of atoms. We select the state-of-the-art work SchNet as a comparison for better illustration. Four representative and prevalent properties (μ, $\varepsilon_{\text{HOMO}}$, U, Cv) are picked in this experiment. The dotted horizontal line in each subplot is the chemical accuracy of each property. Figure 3 shows that MGCN assesses more accurate and stable performance than SchNet. Furthermore, as the number of atoms increases, the advantage of MGCN becomes more apparent due to the multilevel modeling. Our model simplifies the interactions by dividing them into different levels and represent them respectively using the mulitilevel structure and decomposition of molecular quantum interactions. Along this line, our model performs better comparatively when the number of atoms increases. In addition, the significant fluctuations appearing in the front and end of the curves derive from the lack of molecules that contain less than 10 atoms or more than 24 atoms in dataset.

To investigate this further, we construct a control model that blends all levels of interactions in a single embedding rather than construct representation level by level. Taking the prediction of U property for example, the result of this model is not as well as our MGCN with an MAE of 0.03683 on average. It implies the molecular representations modeled by multilevel interaction layers are more robust, which validates the effectiveness of our multilevel modeling.

Generalizability. The potential molecules in chemical space is numerous extra, but the amount of labeled data is quite small (Schütt et al. 2017b). Due to the limitation of the magnitude of existing datasets, the ideal model should be able to perform well even trained with a small amount of

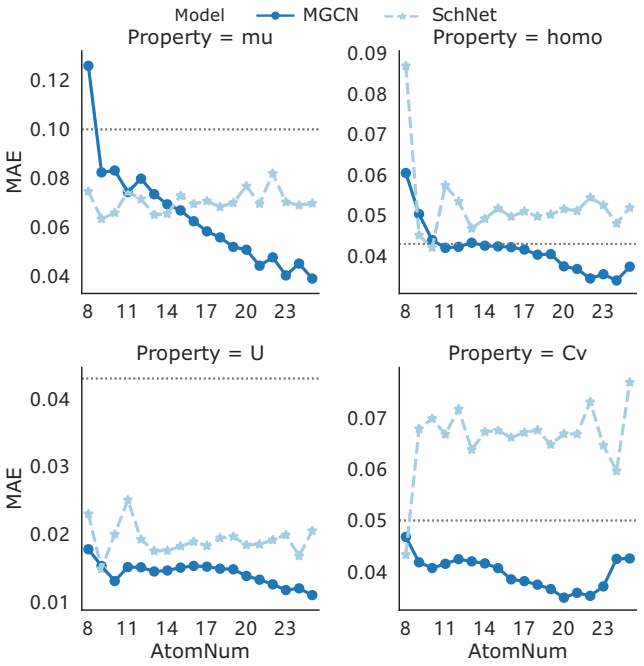

Figure 3: MAE of prediction in different size molecules.

Table 3: Performance comparison in varied size training set

N	SchNet	DTNN	enn-s2s	MGCN
50,000	0.0256	0.0408	0.0249	**0.0229**
100,000	0.0147	0.0364	-	**0.0142**
110,462	0.0134	-	0.0194	**0.0129**

data. Thus, the generalizability is another essential aspect to evaluate these models.

We train our model in three training sets with the different size that consist of 50k, 100k and 110k samples respectively and test them in the same test set that contains 10k molecules. In addition to the three baselines mentioned above (DTNN, SchNet, enn-s2s), the ensemble model of five enn-s2s models is also listed in the comparison. In Table 3, the MGCN gets the lowest MAE in three training sets. Regarding that the readout phase of MGCN and SchNet are similar, the representation of molecular learned by MGCN is more generalizable when the accessible data is smaller thanks to the modeling of multilevel interactions.

Transferability. As mentioned before, the data in existing datasets are unbalanced. For instance, relative large molecules that contain more than 20 atoms account for merely 20.7% of total amount in QM9 and only occupy

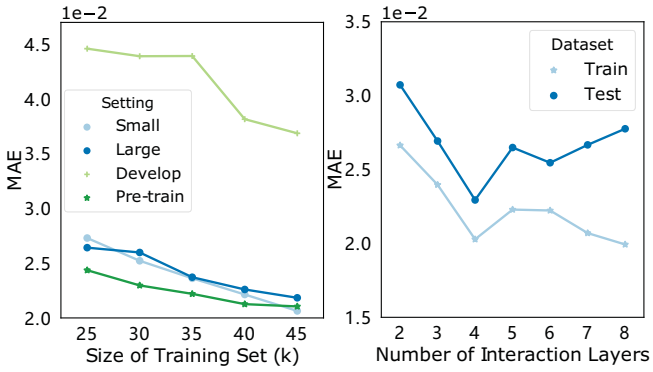

Figure 4: a(left). Performance comparison in the training set with different size. b(right). Predictive performance of models with different number of interaction layers.

8.5% in ANI-1. Therefore, the transferability is quite important to an approach.

We conduct experiments to validate the transferability of MGCN. Specifically, we sample 50k small and 50k large molecules from QM9 respectively. As shown in Figure 4.a, there are four experimental settings. The first two models are trained and tested on the small and large molecules respectively (noted by "Small" and "Large"). The latter two approaches are different ways to transfer the knowledge. The "Develop" denotes the develop model which is trained on the small molecules and applied to the large ones directly. The last model (labeled with "Pre-train") utilizes the embeddings learned on the small molecules as initialization, and then refine itself during the training on the large molecules.

Figure 4.a illustrates the performance of MGCN in four settings. In the first place, The MAE of small molecules is lower relatively due to the higher complexity of large molecules. Secondly, we observe that as we feed in more small data, the MAE of develop model keeps decreasing. The performance is fairly decent because we did not feed any large molecules to this model. The numerical results show that our model is capable to learn knowledge from small data and then transfer them to larger molecules. Thirdly, the pre-trained model outperforms all of other models when the size of training set is small with the universal domain knowledge learned before. This technique helps address the structural shortage of data.

Influence of interaction layers. In DFT, physicists usually use 4 to 5 different empirical symmetry functions for molecular property prediction. Each symmetry could be mapped to the interaction layer in each level. Figure 4.b shows the relationship between the number of interaction layers and the MAE of property U_0. We randomly pick 50k molecules as our training set and test on remaining data. As Figure 4.b illustrates, too many or few interaction layers could cause higher MAE. The network with less than 4 interaction layers does not have enough capacity to learn

the representations of molecules and using the deeper model that contain more than 5 interaction layers will widen the generalization gap. The empirical results indicate that four is the best number of interaction layers which conform to the number of symmetry functions mentioned previously.

Summary. Through the experiments, MGCN shows the superiority of incorporating multilevel modeling of molecular interactions. Moreover, the experimental results prove that our model is generalizable and transferable. Besides, in theory, the time complexity of MGCN is $\mathcal{O}(N^2)$ compared with $\mathcal{O}(N^3)$ of DFT. Experimentally, with the same setting (a single core of a Xeon E5-2660), our model spends 2.4×10^{-2} second predicting the property of one molecule, which is nearly 1.5×10^{5} times faster than DFT.

Conclusion

In this paper, we introduced a Multilevel Graph Convolutional Network (MGCN) for molecular property prediction. The well-designed model utilized the multilevel structure in molecular system to learn the representations of the quantum interactions level by level, and then made prediction with overall interaction representation. The experimental results on two prevalent datasets demonstrated the competency of our approach. Furthermore, our model was proved to be generalizable and transferable.

We believe future research should concentrate efforts on enhancing the generalization of the atom representation because the predictive accuracy is quite high in small samples and it is tough to obtain the dataset of sufficient large molecules.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (Grants No. 61672483, 11774327), and the Science Foundation of Ministry of Education of China & China Mobile (No. MCM20170507). Qi Liu gratefully acknowledges the support of the Young Elite Scientist Sponsorship Program of CAST and the Youth Innovation Promotion Association of CAS (No. 2014299).

References

Becke, A. 2007. *The quantum theory of atoms in molecules: from solid state to DNA and drug design*. John Wiley & Sons.

Behler, J. 2014. Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics: Condensed Matter* 26(18):183001.

Boomsma, W., and Frellsen, J. 2017. Spherical convolutions and their application in molecular modelling. In *NIPS*, 3436–3446.

Bronstein, M. M.; Bruna, J.; LeCun, Y.; Szlam, A.; and Vandergheynst, P. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34(4):18–42.

Broomhead, D. S., and Lowe, D. 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom).

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.

Cubuk, E. D.; Malone, B. D.; Onat, B.; Waterland, A.; and Kaxiras, E. 2017. Representations in neural network based empirical potentials. *The Journal of chemical physics* 147(2):024104.

Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; and von Lilienfeld, O. A. 2017. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of chemical theory and computation* 13(11):5255–5264.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *ICML*.

Goh, G. B.; Siegel, C.; Vishnu, A.; and Hodas, N. O. 2017. Chemnet: A transferable and generalizable deep neural network for small-molecule property prediction. *arXiv preprint arXiv:1712.02734*.

Gómez-Bombarelli, R.; Duvenaud, D. K.; Hernández-Lobato, J. M.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic chemical design using a data-driven continuous representation of molecules. In *ACS central science*.

Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; and Tkatchenko, A. 2015. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters* 6(12):2326–2331.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *CVPR* 770–778.

Hohenberg, P., and Kohn, W. 1964. Inhomogeneous electron gas. *Physical review* 136(3B):B864.

Huang, B., and von Lilienfeld, O. A. 2016. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity.

Huang, Z.; Liu, Q.; Chen, E.; Zhao, H.; Gao, M.; Wei, S.; Su, Y.; and Hu, G. 2017. Question difficulty prediction for reading problems in standard tests. In *AAAI*, 1352–1359.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. *CVPR* 1725–1732.

Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; and Riley, P. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30(8):595–608.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Kohn, W., and Sham, L. J. 1965. Self-consistent equations

including exchange and correlation effects. *Physical review* 140(4A):A1133.

Kollman, P. 1985. Theory of complex molecular interactions: computer graphics, distance geometry, molecular mechanics, and quantum mechanics. *Accounts of Chemical Research* 18(4):105–111.

Lawless, W., and Chandrasekara, R. 2002. Information density functional theory: A quantum approach to intent. In *Proceedings AAAI Fall Conference*.

LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361(10):1995.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *CoRR* abs/1801.07606.

Liu, Q.; Huang, Z.; Huang, Z.; Liu, C.; Chen, E.; Su, Y.; and Hu, G. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1821–1830. ACM.

McDonagh, J. L.; Silva, A. F.; Vincent, M. A.; and Popelier, P. L. 2017. Machine learning of dynamic electron correlation energies from topological atoms. *Journal of chemical theory and computation* 14(1):216–224.

Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; Lilienfeld, A. V.; and Müller, K.-R. 2012. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, 440–448.

Oglic, D.; Garnett, R.; and Gärtner, T. 2017. Active search in intensionally specified structured spaces. In *AAAI*, 2443–2449.

Ramakrishnan, R., and von Lilienfeld, O. A. 2015. Many molecular properties from one kernel in chemical space. *CHIMIA International Journal for Chemistry* 69(4):182–186.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 1.

Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; and Reymond, J.-L. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling* 52 11:2864–75.

Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017a. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NIPS*, 992–1002.

Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; and Tkatchenko, A. 2017b. Quantum-chemical insights from deep tensor neural networks. *Nature communications* 8:13890.

Shang, C.; Liu, Q.; Chen, K.-S.; Sun, J.; Lu, J.; Yi, J.; and Bi, J. 2018. Edge attention-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1802.04944*.

Thouless, D. J. 2014. *The quantum mechanics of many-body systems*. Courier Corporation.

Wang, J., and Hou, T. 2011. Application of molecular dynamics simulations in molecular property prediction ii: diffusion coefficient. *Journal of computational chemistry* 32(16):3505–3519.

Yanai, T.; Tew, D. P.; and Handy, N. C. 2004. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical Physics Letters* 393(1-3):51–57.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X.; et al. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, volume 2, 6.

Zhu, H.; Liu, Q.; Yuan, N. J.; Qin, C.; Li, J.; Zhang, K.; Zhou, G.; Wei, F.; Xu, Y.; and Chen, E. 2018. Xiaoice band: A melody and arrangement generation framework for pop music. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2837–2846. ACM.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.