Molecular Representation Learning by Leveraging Chemical Information

Weibin Li[*], Shanzhuo Zhang[*], Lihang Liu, Zhengjie Huang, Jieqiong Lei,
Xiaomin Fang, Shikun Feng, and Fan Wang

Baidu Inc., China
{liweibin02,zhangshanzhuo}@baidu.com

Abstract. Molecular property prediction is of great importance in AI drug design due to its high experimental efficiency compared with biological experiments. As graph neural networks have achieved great success in many domains, some studies apply graph neural networks to molecular property prediction and regard each molecule as a graph. A molecule's atom is regarded as a node of the graph, while its bond is regarded as an edge of the graph. However, most existing methods simply apply general graph neural networks without considering the domain knowledge. As chemical information is highly related to molecular functions, it is critical for accurate property prediction. Thus, we leverage chemical information to learn molecular representation by integrating molecular fingerprints, i.e., the presence or absence of particular chemical substructures. We compare our proposed method to several strong baselines, and our proposed method significantly surpasses other methods. Up to now, our method ranks first in the Open Graph Benchmark(OGB) leaderboard for ogbg-molhiv.

Keywords: Representation Learning · Graph neural network · Chemical Information.

1 Introduction

The process of drug discovery is complex and full of challenges. Discovering a new drug requires a significant number of biological experiments, which costs hundreds of millions of dollars and tens of years. Molecular property prediction is one of the most critical steps in drug discovery, and the improvement of the prediction accuracy can greatly speed up the process of drug discovery.

Recently, many studies apply machine learning methods to molecular property prediction due to its low experimental cost and high experimental efficiency. Owning to the great success of graph neural networks (GNNs) in many domains, such as recommender systems and social networks, some studies [13] exploit GNNs for molecular property prediction, and each molecule is regarded as a

[*] Authors contribute equally to this work.

graph. More concretely, a molecule's atom is regarded as a node of the graph, while the molecule's bond is regarded as an edge of the graph. Graph neural networks provide accurate predictions by capturing the relations between the atoms in the molecules. However, most of the existing studies directly copy the methods from other domains to molecular property prediction, ignoring the domain knowledge.

Molecular property prediction is highly dependent on chemical information, as the chemical information is always related to molecular functions. To this end, we consider chemical information and combine it with graph neural networks. We design a molecular representation learning task by predicting molecular fingerprints before predicting molecular properties. We adopt two molecular fingerprints: 1) The Molecular ACCess System (MACCS) key [3], which marks substructures of importance (such as functional groups) in determining the macro-chemical properties of a molecule. 2) The extended-connectivity fingerprint (ECFP) [12], which encodes the local structural information of each atom.

We conducted extensive experiments on several datasets to demonstrate the superiority of our proposed method. Up to now, our method ranks first in the Open Graph Benchmark(OGB)[1] leaderboard for ogbg-molhiv.

Our contributions can be summarized as follows:

- We propose a novel molecular representation learning method that takes the chemical information into consideration.
- Extensive experiments show that our proposed method surpasses several strong baselines, and our method ranks first in the OGB leaderboard for ogbg-molhiv.

2 Methodology

We propose a two-stage method. In the first stage, we learn the molecular chemical representation by the graph neural network. In the second stage, we take the molecular representation as input and apply the random forest to predict the molecular property.

2.1 Representation Learning

Graph Neural Networks. A molecule can be regarded as a collection of atoms and the chemical bonds between them, which is naturally suitable to be modelled using graph neural networks (GNNs) [5, 14]. A GNN treats molecules as a graph, where nodes represent atoms and edges represent chemical bonds. By training this GNN, we can obtain a representation vector h_G for the entire graph and a vector h_v for every node $v \in G$.

Modern GNNs follow a neighborhood aggregation approach, where we iteratively update the node's representation by aggregating representations of its

[1] http://ogb.stanford.edu/

neighbors. After k iterations of aggregation, a node's representation captures the structural information within its k-hop network neighborhood. Formally, the k-th layer of a GNN is

$$a_v^{(k)} = \text{AGGREGATE}^{(k)}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \mathcal{N}(v)\right\}\right) \tag{1}$$

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, a_v^{(k)}\right) \tag{2}$$

where $h_v^{(k)}$ is the feature vector of node v at the k-th iteration. We initialize $h_v^{(0)} = X_v$, and $\mathcal{N}(v)$ is a set of nodes adjacent to v.

Unlike Convolutional Neural Networks (CNNs), which are able to take advantage of stacking very deep layers, GCNs suffer from vanishing gradient, over-smoothing, and over-fitting issues when going deeper. To encode the whole molecular structure information, we utilize DeeperGCN[9], a GNN model that can train very deep GCNs.

We use the *SoftMax* function, which has been studied in many machine learning areas[8], as our aggregation function, which is described as:

$$SoftMax\_Agg_{\beta}(\cdot) = \sum_{u \in \mathcal{N}(v)} \frac{\exp\left(\beta \mathbf{m}_{vu}\right)}{\sum_{i \in \mathcal{N}(v)} \exp\left(\beta \mathbf{m}_{vi}\right)} \tag{3}$$

where $\mathbf{m}_{vu} \in \mathbb{R}^D$, is the given message set $\{\mathbf{m}_{vu} | u \in \mathcal{N}(v)\}$. β is a continuous variable called an inverse temperature.

The COMBINE step can be a concatenation operation followed by a linear mapping. In this work, we use the *sum* function. Thus, the AGGREGATE and COMBINE steps are integrated as follows:

$$\mathbf{m}^l = \mathbf{m}_v^{l-1} + \sum_{u \in \mathcal{N}(v)} \frac{\exp\left(\beta \mathbf{m}_{vu}^{l-1}\right)}{\sum_{i \in \mathcal{N}(v)} \exp\left(\beta \mathbf{m}_{vi}^{l-1}\right)} \mathbf{m}_u^{l-1} \tag{4}$$

To further improve the performance, we also integrate the APPNP[6] algorithm into our model. Specifically, we regard the output of the deeperGCN as the input of the APPNP layer. APPNP utilizes the relationship between graph convolutional networks (GCN) and PageRank[11] to derive an improved propagation scheme based on personalized PageRank. More precisely, APPNP achieves linear computational complexity by approximating topic-sensitive PageRank via power iteration. The formulas can be calculated as:

$$\begin{aligned} \boldsymbol{Z}^{(0)} &= \mathbf{M}^L \\ \boldsymbol{Z}^{(k+1)} &= (1-\alpha)\hat{\hat{\boldsymbol{A}}}\boldsymbol{Z}^{(k)} + \alpha \boldsymbol{M}^L \end{aligned} \tag{5}$$

Here, $\mathbf{M}^L \in \mathbb{R}^{N \times D}$, is the node representations produced by the DeeperGCN. $\hat{\hat{\boldsymbol{A}}}$ is the adjacency matrix with added self-loops and normalization.

GNN aggregates node representations from the final iteration to obtain the representation z_G of the entire graph. The grpah pooling function (denoted as

POOL) can be descripted as:

$$g = \mathrm{POOL}\left(z_v^K, v \in G\right). \tag{6}$$

POOL is a graph pooling function, such as summation or a more sophisticated graph-level pooling function.

Therefore, given a set of graphs $\mathcal{G} = \{\mathbf{G}_1, \mathbf{G}_2, ..., \mathbf{G}_Q\}$, we first obtain the node representations of each graph by AGGREGATE and COMBINE steps, and then use POOL function to obtain the graph representation $\{g_1, g_2, ..., g_Q\}$. In this work, we use *mean* function for POOL function.

Chemical Information. As chemical information is highly related to molecular functions, we design a molecular representation learning task by predicting molecular fingerprints. More concretely, we leverage two types of widely used fingerprints: the MACCS Fingerprint and ECFP Fingerprint. Both are generated by the popular open-source cheminformatics package RDKit[2].

The labeling of molecular substructures in MACCS can be considered as a fusion of chemical expert knowledge. MACCS has a total of 166 bits, each of which is marked with a special molecular substructure. For example, the 154th position in MACCS is 1 means that the molecule contains the carbonyl group (C=O). This functional group is susceptible to corrosion by nucleophiles and can undergo nucleophilic addition reactions under certain conditions. By learning the MACCS, our model can establish the correlation between compounds and biologically active substances, therefore obtain a molecular representation with more chemical information.

ECFP is another widely used molecular fingerprint, which is generated by a variant of the Morgan algorithm[10]. The idea of ECFP is to encode the local environmental information of each atom in the molecule. Different flavors of ECFPs may be generated by selecting different maximum diameters of atom neighborhood and/or different lengths of output. In our model, the diameter is set to 4 and length to 2048. Compared with MACCS, ECFP more focuses on the local topology of a molecule and can represent the presence of non-predefined substructures. By concatenating these two fingerprints, we try to achieve a balance between making our model more chemistry-specific and more topology-specific.

2.2 Molecular Property Prediction

Random forest [4] is a tree-based machine learning method, widely used in many fields due to its effectiveness and efficiency. We apply the random forest as a classifier and take the representation learned by the graph neural network to predict the molecular property.

[2] https://www.rdkit.org/

3 Experiments

3.1 Experimental Settings

In this work, the main GNN model applied to learn the molecular representation is DeeperGCN[9] and APPNP[6]. We select the following hyper-parameters in our experiments: 7 layers DeeperGCN, 256-dimensional hidden size and average graph pooling for POOL function. We use alpha=0.2 and k_hop=5 for APPNP layer.

In the first training stage, Adam optimizer with a learning rate of 0.001 is adopted and we train our model for 50 epochs.

In the second training stage, we use a random forest classifier to predict the molecular property by taking the molecular representation as input. We run 10 experiments with different random seeds for each property prediction task.

3.2 Baseline Methods

We compare our models above with a number of state-of-the-art baselines for graph property prediction.

- **Random Forest[4].**Random forest is a tree-based classifier. The essence is to build multiple trees in randomly selected subspaces of the feature space.
- **DGN[1].** Directional Graph Network exploits vector flows over graphs and asymmetric aggregation functions.
- **DeeperGCN[9]+FLAG[7].** DeeperGCN defines differentiable generalized aggregation functions to unify different message aggregation operations. Free Large-scale Adversarial Augmentation on Graphs(FLAG) iteratively augments node features with gradient-based adversarial perturbations.
- **PNA[2].** PNA combines multiple aggregators with degree-scalers.

3.3 Experimental Results

We compare our proposed method with several strong baselines, as shown in Table 1. Our proposed method outperforms other methods as it learns the molecular representation from the chemical information.

Table 1. Performance on dataset Hiv

Method	Test AUC	Valid AUC
Ours	**0.8232 ± 0.0047**	**0.8331 ± 0.0054**
Random Forest[4]	0.8060 ± 0.0010	0.8420 ± 0.0030
DGN[1]	0.7970 ± 0.0091	0.8470 ± 0.0047
DeeperGCN+FLAG[9, 7]	0.7942 ± 0.0120	0.8425 ± 0.0061
PNA[2]	0.7905 ± 0.0132	0.8519 ± 0.0099

4 Conclusion

We propose a two-stage method to address the problem of molecular property prediction. In the first stage, we integrate chemical information with a graph neural network and learn the molecular representation. In the second stage, we apply the random forest to predict the molecular property by taking the molecular representation as input. This method achieves excellent performance in the dataset Hiv.

References

1. Beaini, D., Passaro, S., Létourneau, V., Hamilton, W.L., Corso, G., Liò, P.: Directional graph networks. arXiv preprint arXiv:2010.02863 (2020)
2. Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P.: Principal neighbourhood aggregation for graph nets. arXiv preprint arXiv:2004.05718 (2020)
3. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of MDL Keys for Use in Drug Discovery. Journal of Chemical Information and Computer Sciences **42**(6), 1273–1280 (Nov 2002)
4. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
5. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., Leskovec, J.: Strategies for pre-training graph neural networks. arXiv preprint arXiv:1905.12265 (2019)
6. Klicpera, J., Bojchevski, A., Günnemann, S.: Predict then propagate: Graph neural networks meet personalized pagerank. arXiv preprint arXiv:1810.05997 (2018)
7. Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., Goldstein, T.: Flag: Adversarial data augmentation for graph neural networks. arXiv preprint arXiv:2010.09891 (2020)
8. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
9. Li, G., Xiong, C., Thabet, A., Ghanem, B.: Deepergcn: All you need to train deeper gcns. arXiv preprint arXiv:2006.07739 (2020)
10. Morgan, H.L.: The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. Journal of Chemical Documentation **5**(2), 107–113 (May 1965), https://doi.org/10.1021/c160017a018
11. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
12. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling **50**(5), 742–754 (2010)
13. Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. Chemical science **9**(2), 513–530 (2018)
14. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)