Recognition of Molecular Characteristics Considering Graph Structure

Sho Ishida, Tomo Miyazaki, Yoshihiro Sugaya, Shinichiro Omachi
Tohoku University

Abstract—**In recent years, research on recognition of molecular properties using neural networks has been widely conducted. They are focusing on the graph structure to extract features from molecular. In this study, we aim to recognize molecular properties with high accuracy by combining existing method.**

I. INTRODUCTION

Analyzing molecules is a fundamental step in chemical research such as drug discovery and biological reaction elucidation, and plays an important role. At present, however, there are an infinite number of molecules, so research on recognition of molecular properties by machine learning is actively conducted. Here, since the characteristics of molecules are closely related to their structural features. A technique often used to obtain structural features is to treat molecules as graphs and obtain graph features. In this study, we considered a method to improve the recognition accuracy with reference to the existing method to obtain the characteristics of the graph.

II. RELATED WORK

A. Feature design

When performing molecular recognition, first replace molecules with features. Based on this feature, the computer learns and recognizes the characteristics, but the recognition accuracy varies greatly depending on the feature quantity design method. In this study, two feature extraction methods focusing on the graph structure of molecules were used. Both methods treat a compound as a graph by using atoms in the compound as nodes and bonds as edges.

B. Graph Convolutional Network (GCN)[1]

GCN is a technique for obtaining graph features by assigning feature vectors to graph nodes and using the Convolution layer and Pooling layer to update the node features. In the Convolution layer, the attention node and the adjacent node are weighted, and the feature of the node is updated by adding them and applying the activation function. In the pooling layer, nodes are updated by taking the average or maximum value of the attention nodes and adjacent nodes. Nodes updated in this way are finally combined into one feature by taking the average of all nodes.

C. Extended Connectivity Fingerprint (ECFP)[2]

ECFP is a method of expressing the presence or absence of structural features of graphs using 0/1 vectors. First, the structural features in the graph are listed in values by a hash function. A fixed-length 0 vector is prepared in advance, and 1 is set to the bit corresponding to the value obtained by dividing the hash value by the fixed length. In this way, a fixed-length 0/1 vector representing the presence or absence of structural features is generated.

III. PROPOSED METHOD

In this study, a new feature amount was defined by combining two feature amount design methods, GCN and ECFP [2]. In GCN, graph features are obtained by convolving nodes in a graph using edge information. ECFP is a method of representing the presence or absence of structural features in a graph with a 0/1 vector. With GCN alone, information on atoms and neighboring atoms is regarded as important, and it is thought that structural features are missing. Therefore, as shown in Figure 1, it was considered that high-precision characteristic recognition can be realized by adding to the ECFP feature quantity GCN which is a structural feature. In GCN, features were extracted by applying two layers of convolution and pooling. For ECFP features, we added a fully connected layer and combined features extracted from ECFP features by learning. The combined feature was recognized by one fully connected layer.

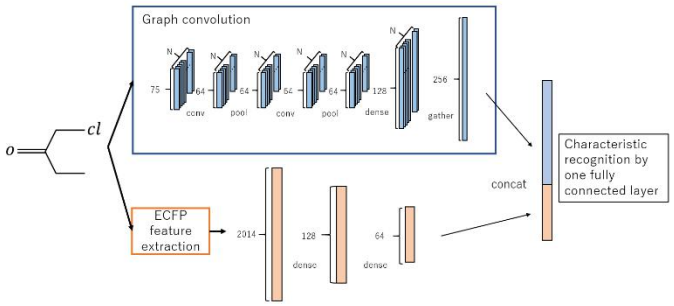

Figure1. Graph features are extracted by GCN and ECFP and combined. GCN uses two layers of convolution and pooling. ECFP obtains a0/1 vector of fixed length 2014. It is then weighted by two fully connected layers and connected to GCN feature, then recognized by one fully connected layer.

IV. EXPERIMENT

A. Dataset

In the experiment, Tox21, ToxCast, ClinTox, BBBP, BACE, HIV, SIDER were used as Recognition data and QM7, QM8, QM9, ESOL, PDBbind, and Lipophilicity were used as Classification data. Each is a data set for evaluating molecular properties such as toxicity and solubility, and it stores

molecular data described in smiles notation. Details of the dataset are shown in Table 1.

Category	Dataset	Task		Compounds
Quantum mechanics	QM7	1	Regression	7165
	QM8	12	Regression	21786
	QM9	12	Regression	133885
Physical chemistry	ESOL	1	Regression	1128
	Lipophilicity	1	Regression	4200
Biophysics	PDBbind	1	Regression	11908
	BACE	1	Classification	1522
Physiology	BBBP	1	Classification	2053
	Tox21	12	Classification	8014
	ToxCast	617	Classification	8615
	SIDER	27	Classification	1427
	ClinTox	2	Classification	1491

TABLE I. DATASET DETAIL

B. Evaluation method

ROC-AUC score was used for classification data as an evaluation method. In regression data, MAE score was used in QM7, QM8, and QM9, and in other data, RMSE score was used. The ROC-AUC score is represented by the area covered by a curve plotted with false positives on horizontal axis and true positives on the vertical axis. RMSE score and MAE score are calculated by the following equations:

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_i \left(\boldsymbol{y_{obs,i}} - \boldsymbol{y_{pred,i}}\right)^2}{\boldsymbol{n}}} \#(\mathbf{1})$$

$$\boldsymbol{MAE} = \frac{\sum_i \left|\boldsymbol{y_{obs,i}} - \boldsymbol{y_{pred,i}}\right|}{\boldsymbol{n}} \#(\mathbf{2})$$

Where y_{obs} is the observed value, y_{pred} is the predicted value, and n is the number of data. The ROC-AUC score is better as the value is larger, and the MAE and RMSE scores are better as the value is smaller.

V. RESULT

Tables 1 and 2 show the results of comparison between the score of ECFP and Graph Convolution and the score of the proposed method. The score of the proposed method rose on average for both classification data and regression data. It can be considered that the combination of the two methods compensated for the lack of information. The overall ECFP score was poor, but adding the fully connected layer reduced the effect of ECFP on the combined features, so the proposed method gave good results. However, in this experiment, it was not possible to confirm whether the intended structural features could be supplemented by ECFP.

TABLE II. ROC-AUC SCORE

	Tox21	ToxCast	ClinTox	BBBP	BACE	HIV	SIDER
ECFP	0.775	0.609	0.705	0.874	0.856	0.713	0.582
Graph Conv	0.807	0.718	0.776	0.713	0.834	0.738	0.605
Proposed	**0.820**	**0.720**	**0.838**	**0.719**	**0.850**	**0.758**	**0.602**

TABLE III. MAE SCORE, RMSE SCORE

	MAE			RMSE		
	QM7	QM8	QM9	ESOL	PDBbind	Lipophilicity
ECFP	239	0.0579	47.9	1.61	2.01	0.978
Graph Conv	122	0.0288	22.9	1.04	2.05	0.809
Proposed	**118**	**0.0279**	**22.9**	**1.01**	**2.06**	**0.772**

VI. CONCLUSION

A Graph ECFP feature by combining the features of Graph convolution and ECFP is proposed, and the recognition accuracy is improved. Future work include the consideration of molecular features such as the ring structure, enantiomers, and edges.

References

[1] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande, Low Data Drug Discovery with One-shot Learning 2016

[2] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints.j. Chem. Inf. Model.2010, 50(5): 742-754

[3] Zhenqin Wu, Bharath Ramsundar, Evan N. Feiberg, Joseph gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing and Vijay Pnde, MoleculeNet: a benchmark for molecular machine learning