VIRTUALMAN www.virtualman.ai contact@virtualman.ai

A new explainable graph convolution network based on discrete method: using water solubility as an example

Hsin Liu [a,c], Cheng-Wei Lee [a], Bo-Han Su [b], and Yufeng J. Tseng [b,c,d,e]

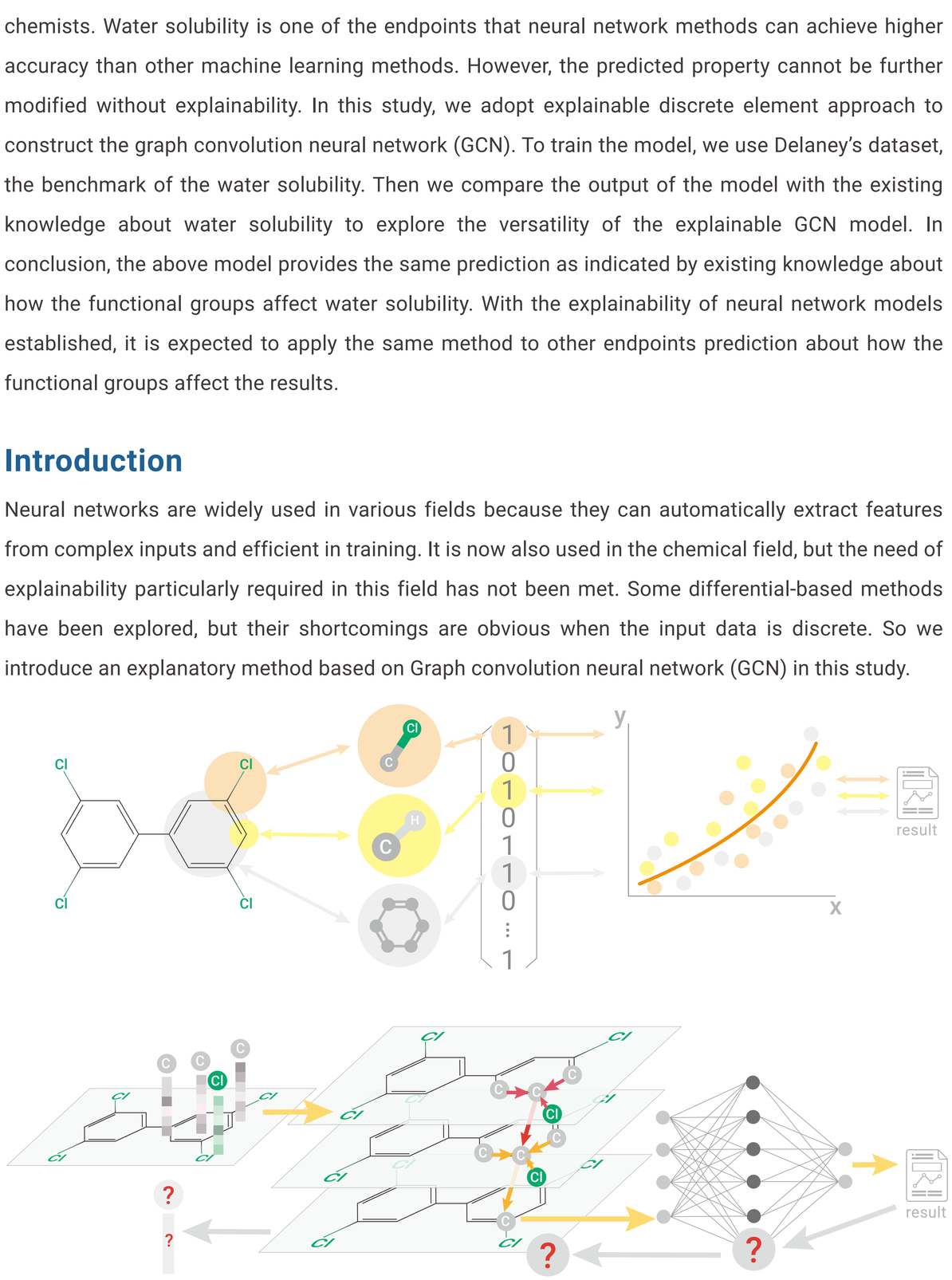
a. Virtual Man Inc. 3F., No. 196, Sec. 2, Fuxing S. Road, Da'an Dist., Taipei 106, Taiwan
b. Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.
c. Graduate Institute of Biomedical Engineering and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.
d. The Metabolomics Core Laboratory, Center of Genomic Medicine, National Taiwan University, No. 1 Sec. 1, Jen Ai Road, 106, Taiwan
e. Drug Research Center, School of Pharmacy, College of Medicine, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.

Abstract

Over the past few years, the effectiveness of neural network methods has been approved in several fields, including image recognition, natural language processing, and chemical properties predicting. While the neural network provides higher accuracy, its lack of explainability tends to limit the use for chemists. Water solubility is one of the endpoints that neural network methods can achieve higher accuracy than other machine learning methods. However, the predicted property cannot be further modified without explainability. In this study, we adopt explainable discrete element approach to construct the graph convolution neural network (GCN). To train the model, we use Delaney's dataset, the benchmark of the water solubility. Then we compare the output of the model with the existing knowledge about water solubility to explore the versatility of the explainable GCN model. In conclusion, the above model provides the same prediction as indicated by existing knowledge about how the functional groups affect water solubility. With the explainability of neural network models established, it is expected to apply the same method to other endpoints prediction about how the functional groups affect the results.

Introduction

Neural networks are widely used in various fields because they can automatically extract features from complex inputs and efficient in training. It is now also used in the chemical field, but the need of explainability particularly required in this field has not been met. Some differential-based methods have been explored, but their shortcomings are obvious when the input data is discrete. So we introduce an explanatory method based on Graph convolution neural network (GCN) in this study.



Material & Methods

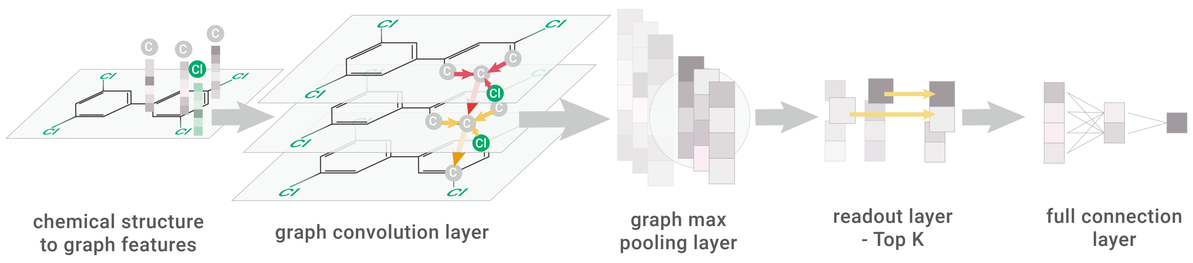
Dataset Retrieval & Curation

Delaney small dataset (Delaney et al.1) is the most commonly used water-soluble data containing 1128 records with the aqueous solubility values as log mol/L at 25 °C. This dataset is widely used in several related works about solubility model construction.

ChemIDPlus database was built from more than 100 data sources and mainly matched to the compounds cited in National Library of Medicine (NLM). Our study retrieved and cleaned the data from the above database, which included 1967 compounds.

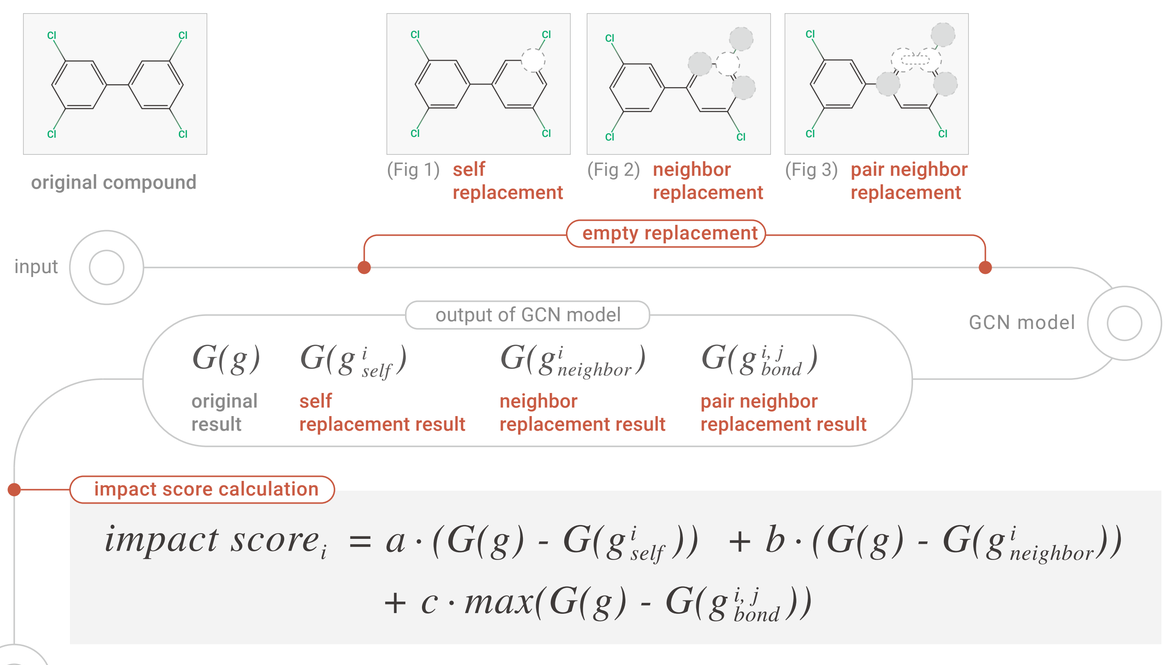
Methods

Graph Convolutional Neural Network (GCN)[1] is used in this study as the prediction method. The molecular structure is transformed into a graph structure as model input, and each atom is encoded as 77 features.



The model structure 2 graph convolution layer, 1 max pooling layer, 1 top K layer, 5 full connection layer

Discrete Explainable Method is based on GCN. The features of GCN node encode with atom features are binary. For each feature, 1 represents this feature and 0 represents no feature. The GCN model can be denoted as a function G, and the original input can be denoted as g, so the output of the model can be denoted as $G(g)$. The self replacement only replaces the self node i to an empty node denoted as g^{i}_{self} (Fig 1), the neighbor replacement replaces not only the self node i but also all the nodes connected to the node denoted as $g^{i}_{neighbor}$ (Fig 2), and the pair neighbor replacement replaces the 2 nodes i and j with all the neighbors connected by one edge denoted as $g^{i,j}_{bond}$ (Fig 3). We define the node impact score in each replacement as $G(g_{replacement}) - G(g)$. Because the $G(g^{i,j}_{bond})$ should be distributed to each connected node, so we define the impact score of node i for this case as $max(G(g) - G(g^{i,j}_{bond}))$ where j is the nodes connected to i. The three factors produced by above methods will correspond to three coefficients, thus the final impact score of node i is:



$$impact\ score_i = a \cdot (G(g) - G(g^{i}_{self})) + b \cdot (G(g) - G(g^{i}_{neighbor})) + c \cdot max(G(g) - G(g^{i,j}_{bond}))$$

Results

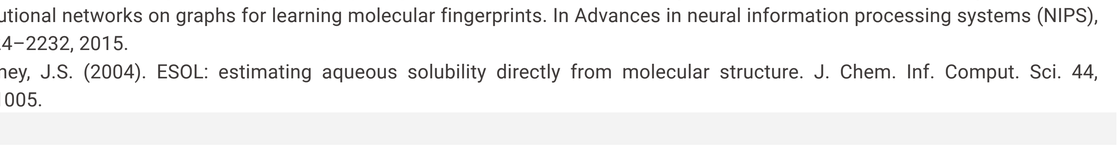
Pearson Correlation Coefficient (PCC)

The model perform by testing dataset indicated by PCC with Delaney small dataset is above 0.95 and with larger Chemidplus dataset the performance by PCC is 0.89.

Delaney small dataset	dataset	Chemidplus
0.98	PCC	0.89
1128	data size	1967

Explainable result

We selected the Chemidplus dataset to train the model because of its large size. The result is as the following, the values shown here are only those above average.



Conclusion

- Numeric value-based dataset also can be used to train the GCN explainable neural network.
- Regarding water solubility, our explainable model proposed the same substructure as indicated by literature review.
- Auto feature extraction works well in GCN neural network training. No features converted from human knowledge were adopted in this study.
- This model can help researchers discover the problematic substructure for other endpoints numeric value-based dataset.

References

1. David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan´ Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In Advances in neural information processing systems (NIPS), pp. 2224–2232, 2015.
2. Delaney, J.S. (2004). ESOL: estimating aqueous solubility directly from molecular structure. J. Chem. Inf. Comput. Sci. 44, 1000–1005.