Computers and Chemical Engineering 179 (2023) 108403

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Graph neural networks with molecular segmentation for property prediction and structure–property relationship discovery

Zhudan Chen [a], Dazi Li [a,*], Minghui Liu [b], Jun Liu [b]

[a] Institute of Automation, Beijing University of Chemical Technology, Beijing, 100029, China
[b] State Key Laboratory of Organic-Inorganic Composites, Beijing University of Chemical Technology, Beijing, 100029, China

ARTICLE INFO

Keywords:
Graph neural networks
Molecular segmentation
Property prediction
Structure–property relationship

ABSTRACT

Graph neural networks (GNNs) have been widely used for predicting properties and discovering structure–property relationships in chemistry and drug discovery. However, current GNNs treat all atoms as independent of each other, neglecting the crucial role of functional groups in molecules. In this work, graph neural networks based on molecular segmentation are proposed. An unsupervised segmentation method is proposed to partition molecular graph data into multiple functional group-based clusters. Segmentation message passing neural network is proposed to learn function groups, which generate embeddings in and between molecule clusters. To explain structure–property relationships, we propose a new explainer to identify substructures that are more compatible with the chemical principles analysis. Our approach attempts to train and explain GNNs more rationally by segmenting molecules. Experimental results show that our approach achieves more efficient and rational performance prediction and explanation on chemical and drug datasets.

1. Introduction

Graph neural networks (GNNs) (Ruiz et al., 2021; Rong et al., 2020; Gilmer et al., 2017; Veličković et al., 2017; Kipf and Welling, 2016) have shown excellent processing performance for graph-structured data and have been applied to various fields (Alon and Yahav, 2020), including knowledge graphs (Ye et al., 2022), social networks (Fan et al., 2022), complex systems (Shi et al., 2022), and molecules (Fradkin et al., 2022; Song et al., 2020). This relies on the fact that Zhou et al. (2020) and Scarselli et al. (2008) graph networks are designed to represent the vertices of a graph as low-dimensional vectors by preserving information about the network topology and node content of the graph for processing using simple machine learning algorithms, such as fully connected neural networks. For molecules, atoms are seen as the nodes of graphs, and the chemical bonds formed between atoms make up the topology of graphs. The analysis of small molecule data in chemistry or drugs using GNNs has become popular for scientists to study. Recently, several GNNs models have been reported in the domains of chemical (Rittig et al., 2023; Fung et al., 2021) and drug science (Fradkin et al., 2022; Yang et al., 2022) for property prediction (Schweidtmann et al., 2023; Aldeghi and Coley, 2022; Wieder et al., 2020; Zhang et al., 2022a) and structure–property relationship discovery (Amamoto, 2022; Yuan et al., 2022; Luo et al., 2020).

GNNs analysis of molecular structures aims to predict chemical or pharmaceutical properties and discover structure–property relationships. However, the current GNNs algorithms (Wu et al., 2021) treat all atoms as independent of each other and ignore the role of functional groups in molecules. It is well known that functional groups are the atoms or atomic groups that determine the chemical properties of organic compounds. To solve this problem, some scholars (Li et al., 2020) have assisted GNNs to learn molecular structures by manually segmenting structures such as functional groups in molecules, but this requires thousands of molecules to be delineated manually and the segmentation models constructed are black-box models, and the results are not completely accurate. Some other scholars (Rong et al., 2020; Fradkin et al., 2022; Cheung et al., 2019), based on the original GNNs, add global structure information or make changes to the step size of selected neighbors (Nt and Maehara, 2019), etc. to improve the learning ability of the GNNs for molecular structures and improve the performance of the model. Since neural networks belong to black box models, although the prediction performance is improved, these works do not focus on the structure–property relationship explanation of molecules. Benefit from the existing explainability methods for CNNs, Some techniques, including Gradient-weighted Class Activation Mapping (Grad-CAM), the Monte Carlo Tree Search (MCTS) approach, Deep neural networks (DNNs) and Excitation Back-Propagation (EB), are developed new versions for GNNs (Pope et al., 2018; Zhang et al., 2022b; Wen et al., 2022). From the experimental data of graph explanation (Ying et al., 2019; Yuan et al., 2022; Lin et al., 2022; Luo et al.,

* Corresponding author.
E-mail addresses: lidz@mail.buct.edu.cn (D. Li), liujun@mail.buct.edu.cn (J. Liu).

https://doi.org/10.1016/j.compchemeng.2023.108403
Received 11 March 2023; Received in revised form 10 May 2023; Accepted 3 September 2023
Available online 6 September 2023

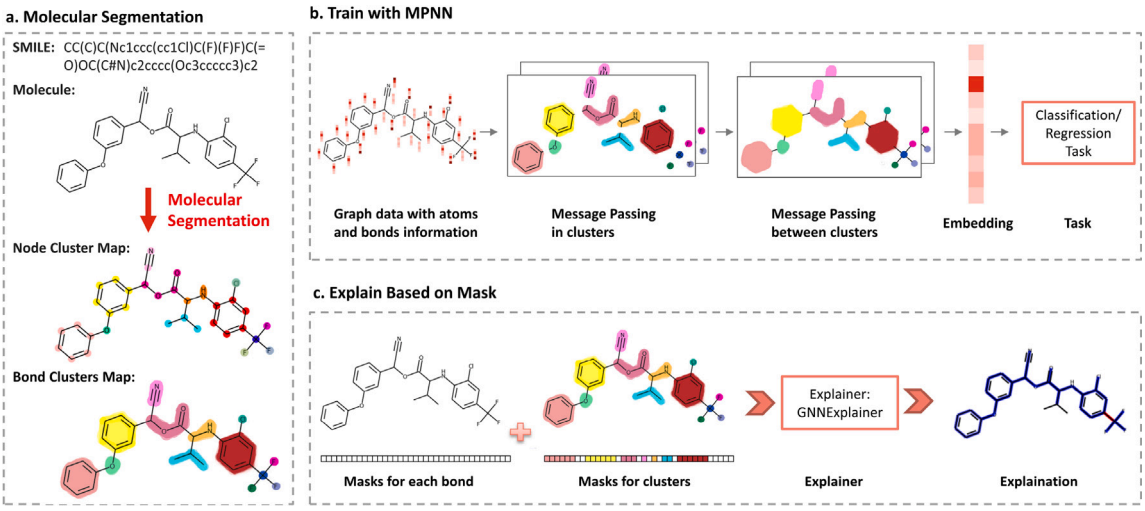

Fig. 1. Workflow of graph neural network method based on molecular segmentation method.

2020; Baldassarre and Azizpour, 2019) as for now, it can be seen that the graph explanation algorithm explains the molecular structure in a discrete way. SubgraphX (Yuan et al., 2021) is designed to obtain connected subgraph explanations, but some of results may not reasonable from molecule insight. The results obtained by existing methods need to be interpreted by professionals or compared by clustering to obtain the relationship between performance and functional group structure, as the explanation methods are not directly relevant to the functional group structure.

When analyzing the effect of molecular structure on performance, the ideal result is that functional groups or specific substructures have a positive or negative effect on the target properties (Blondel, 2003; Wardle and Zackrisson, 2005; Yasir et al., 2022). The current qualification of functional groups or specific substructures is mostly based on common chemical knowledge or the experience of research scholars, so it is not realistic to applied on large amounts of molecules. MolAICal (Shearer et al., 2022) software can partition molecular structures based on the rotable bonds in molecules (Bai et al., 2022; Tian et al., 2022), but the software has significant limitations on the input of molecular structures and is not friendly to databases with diverse molecular structures. An iterative marching algorithm (Ertl, 2017) identify functional groups in a molecule through its atoms is described. The downside is that this algorithm discards the atoms that do not form the functional group structure, leaving the molecule without global integrity. Therefore, it is challenging and necessary to implement molecular segmentation that is mechanistically compatible and can be applied to many databases.

In this paper, a GNN based on molecular segmentation is proposed for molecular property prediction and explanation. Firstly, an unsupervised learning method is proposed to segment molecular graph data into multiple functional group-based categories. To learn the features of functional group clusters, a segmentation message passing neural network is proposed, which generates feature information sequences within and between functional groups after molecular segmentation. To explain the relationship between properties and functional group structures, a new graph explainer is proposed to identify substructures that are more compatible with chemical principle analysis. The flowchart of the proposed method and contains three parts: molecular segmentation method, segmentation-based graph information transfer network, and segmentation-based graph explanation. The work included in this chapter demonstrates the first attempt to train and explain GNNs more rationally by segmenting molecules based on functional groups.

2. Methods

Our methods consists of three part: (a) molecular segmentation method, which segment a whole molecule into several clusters following (b) MPNN; (c) explain. The architecture of our methods is shown in Fig. 1.

2.1. Molecular segmentation method

Molecular segmentation method is used to segment one molecule into different clusters. Cluster maps of atoms and bonds are used in GNNs for property prediction and explanation. Our goal is to keep the molecular fragments as small as possible and preserve the functional fragments, which is helpful for GNNs to learn atomic and substructure information in molecules.

As shown in Fig. 2, four steps are defined to cluster a molecule based on atom type and atom neighbors. Before segmentation, all atoms are in an unmarked state. The first step aims to find all the rings, for example benzene rings and pyridine rings. All unmarked atoms in the same ring are divided to the same cluster. All atoms belong to halogen family are marked in the second step. Each of the halogen atoms is clustered separately because of its unique properties. On the other hand, some atoms, like O and N, may form function fragments, like '-COO-' and '-NO_2', which also have impact on properties. Accordingly, non-carbon atoms are selected in the third step. If the selected atom and its neighbors are all unmarked, a new cluster is set for the atom and its neighbors. If not, the non-carbon atom and its unmarked neighbors would join that cluster that marked atom belongs to. Significantly, only neighbors not in ring structures are considered in the third and forth steps. In the forth step, the remaining atoms are clustered together with their unmarked neighbors. More examples from Solubility dataset are shown in Fig. 3.

The results of the molecular segmentation can be formulated as $\mathbf{C} = (c_1, c_2, \ldots, c_n)$, where $\mathbf{C} \in \mathbb{R}^n$ donates cluster IDs of n nodes, and the i_{th} atom cluster ID is c_i. Atoms with same ID are in the same cluster. Furthermore,with atoms as nodes and chemical bonds as the edges, the molecules are transformed into graph data. Adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$,which is a description of the integrity of the nodes in the graph, of the raw molecule can be developed into two new adjacency matrix $\mathbf{A}_{in}$ and $\mathbf{A}_{out}$ by the following equations:

$$\mathbf{A}_{in} = \begin{cases} a_{ij}, & \text{if } c_i = c_j \\ 0, & \text{others,} \end{cases} \tag{1}$$

$$\mathbf{A}_{out} = \begin{cases} a_{ij}, & \text{if } c_i \neq c_j \\ 0, & \text{others,} \end{cases} \tag{2}$$

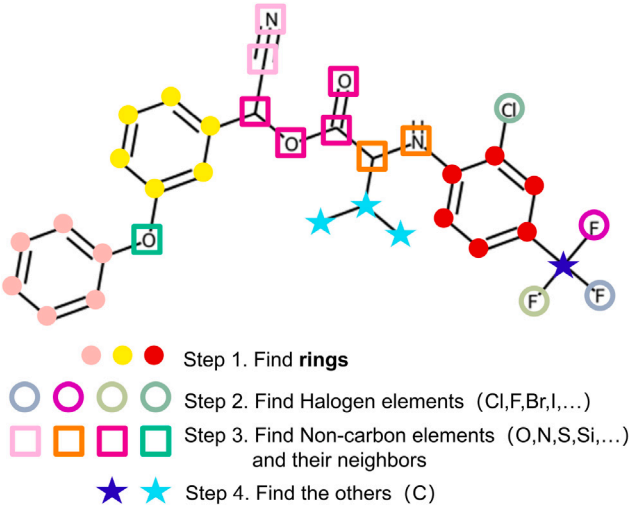

Fig. 2. Schematic diagram of molecular segmentation method steps.

where a_{ij} is the value of row i and column j in $\mathbf{A}$. If there is a bond between the i_{th} atom and the j_{th} atom, a_{ij} is 1, otherwise, a_{ij} is 0. $\mathbf{A}_{in}$ and $\mathbf{A}_{out}$ record the connection situation of molecule segmentation results.

2.2. SMPNN

MPNN (Gilmer et al., 2017) learns features from the raw molecule graph $\mathbf{G}$, containing node feature $\mathbf{X}$ and edge set $\mathbf{E}$, through the popular message-passing mechanism. During message-passing iterations, the representation $\mathbf{h}_u$ of node u aggregates representations from neighbors $\mathcal{N}(u)$ of node u with itself. After l iterations, the representation is updated according to:

$$\mathbf{h}_u^l = \theta_u^{l-1}\mathbf{h}_u^{l-1} + \sum_{w\in\mathcal{N}(u)} \theta_w^{l-1}\cdot\mathbf{h}_w^{l-1} \tag{3}$$

where $\mathbf{h}_u^l$ denotes the representation of node u at iteration l, and θ_u^{l-1}, θ_w^{l-1} are parameters to be trained.

In MPNN feature extraction of graph structure data of molecules, each atom is viewed as a separate node and no attention is paid to the structure of functional groups in the molecule. Segmentation message passing neural network (SMPNN) is proposed to combine the molecular segmentation results with MPNN to produce a new segmentation message passing mechanism. In MPNN, the neighbors $\mathcal{N}(u)$ of node u are generated from the adjacency matrix $\mathbf{A}$. In other words, the message-passing region is the entire original molecular graph topology. In order to make the graph network focus more on the feature information of functional groups, the molecular partitioned categories are used to reduce the confusion of information between unused categories by changing the learning method of the network so that it first learns features within functional groups and later integrates the global structural information. In summary, SMPNN extracts and retains segmentation features by first passing information within molecular segmentation categories and then between categories. To collect features from one segmentation category in the molecule, the message passing region should be narrowed down to each small category topology. The learning can then be continued by considering each category as a new large node and using the connections between all categories to generate feature information for the entire graph. In which, the graph structure information is passed within and between categories, neighbors $\mathcal{N}(u)$ are generated from $\mathbf{A}_{in}$ and $\mathbf{A}_{out}$ when messages are passed within and between clusters, separately.

The flowchart of SMPNN is depicted in Fig. 1(a). The input of SMPNN is graph structure data with node and edge information. First, message passing is performed within the category, which is represented in the figure as the fusion of information only in the color block structure; in other words, nodes select neighboring nodes only in the same category during message passing. After N layers of intra-category message passing, where N is a hyperparameter, the ideal situation is that the information on the node has learned the information of the functional group in which it is located. Then, the global graph topology is learned using the connections between categories. At this point the nodes select neighboring nodes, only the nodes that are in different categories, and with the help of the connection between the two categories, the information of the two functional groups is fused. The formulas for message passing are both Eqs. (3), the difference is that in them the neighbor nodes are selected according to the segmentation of the category. The improvement in the way SMPNN selects neighbor nodes also reduces the number of parameters to be trained in the neural network. Specifically, in Eq. (3), the number of parameters a that need to be trained is the same as the number of neighbors of the nodes. For SMPNN and MPNN with the same number of layers, the MPNN has to train the weights of all neighbor nodes based on the adjacency matrix $\mathbf{A}$, while the SMPNN needs to train the weights of some neighbor nodes based on the new adjacency matrix $\mathbf{A}_{in}$ and $\mathbf{A}_{out}$. Therefore, SMPNN has fewer parameters and lower complexity under the same number of layers.

After segmentation-message-passing, the readout function R is applied to compute an embedding vector for the whole graph:

$$Embedding = R\left(\{\mathbf{h}_u | u\in\mathbf{G}\}\right) \tag{4}$$

where $\mathbf{h}_u$ donates the feature vector of node u after segmentation-message-passing, and $Embedding$ donates the final embedding for the graph. The readout function R obtains the feature representation of the entire graph by aggregating node features. Specifically, R can choose operations such as global maximum pooling, global average pooling, etc. In this work, R is chosen as global average pooling function, and the function can be shown as:

$$R\left(\{\mathbf{h}_u | u\in\mathbf{G}\}\right) = \sigma\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\right) \tag{5}$$

where σ donate a sigmoid function.

With $Embedding$ as ition nput, a fully-connected neural network $f_{nn}(\cdot)$ is trained to achieve property classification or regression as following:

$$Output = f_{nn}\left(Embedding\right) \tag{6}$$

where $Output$ is the classification or regression prediction results for molecule $\mathbf{G}$.

2.3. Explainer based molecular segmentation

The results of molecular segmentation can provide not only the division of regions for message passing in graph networks, but also new definitions for the sequence of weights for graph explanation. In the original graph explainer (Ying et al., 2019), the important edges in the molecule are mostly scattered in the graph structure, and it is difficult to learn the connected substructure. The reason is that the size of the sequence of weights representing the explanation performance is included in the loss function in order for the algorithm to converge when graph explanation is performed. Because information is fused with information from neighboring nodes when passing information in the graph, neighboring nodes with particularly similar feature information after passing through multiple graph network layers can include information about that structure in the final embedding by including only one of them or one edge in a structure. This does not facilitate graph explanation, because the complex structure of the molecule with only one edge does not confirm to which functional group or combination of functional groups the important substructure belongs.

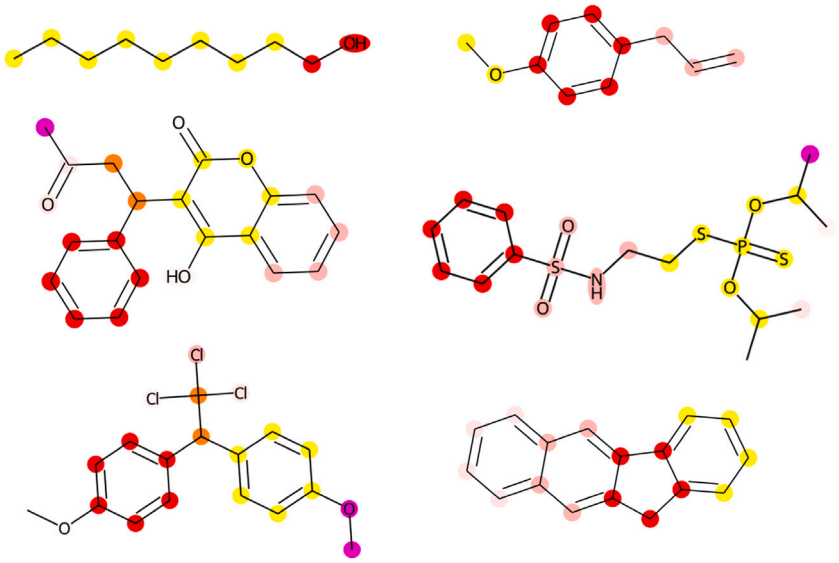

Fig. 3. Schematic diagram of the results of molecular segmentation methods. In each molecule, atoms of the same color belong to the same category.

To explain the predictions of SMPNN, a general approach used in most explain methods is directly explaining the importance of graph edges. Edge mask $\mathbf{M} \in \mathbb{R}^e$ is defined to indicate the probability of each edge appearing in the graph, and the probability values that represent the importance of edges are float type number between 0 and 1. Given a trained GNN model $\mathcal{F}(\cdot)$ and an input graph $\mathbf{G}$, edge mask $\mathbf{M}$ is trained by comparing the outputs of GNN model when the inputs are $\mathbf{G}$ and $\mathbf{G}$ with edge masks. However, this approach cannot provide a connected explanation. In this work, the molecular segmentation provides a new perspective of explainer. The results of molecular segmentation define a new edge mask $\mathbf{M}_c \in \mathbb{R}^e$, in which edges belong to one substructure share the same weight. The proposed explainer based molecular segmentation is illustrated in Fig. 1(c). The goal of our explainer is to figure out one or more important function segmentation clusters for the prediction *Output*. Accordingly, in the new graph $\mathbf{G}_c$, edges in the same cluster share one mask.

In $\mathbf{M}$ from the original graph explanation algorithm, each edge is independent, and e is the number of edges. In $\mathbf{M}_c$ from explainer based on molecule segmentation, the definitions of the weights corresponding to edges in the same category are identical and are also identical in the updating process of the weights. The edges of the same category in the weight masks share one weight, which not only reduces the number of parameters to be trained, but also maintains the integrity in the numerator to a certain extent. It is worth mentioning that the weights of the edges between categories are kept in the weight sequence, which first of all ensures the structural integrity, while the explanation of the graph structure, the result may be a certain functional group or the overall structure.

The parameter $\mathbf{M}_c$ in our explainer is trained by the loss function, inspired from the GNNExplainer, as following:

$$loss = \text{Distance}(\mathcal{F}(\mathbf{G}), \mathcal{F}(\mathbf{G}_c)) + \text{Sum}(\mathbf{M}_c) + \text{Discrete}(\mathbf{M}_c) \tag{7}$$

where Distance$(\cdot)$ measures the distance of $\mathcal{F}(\mathbf{G})$ and $\mathcal{F}(\mathbf{G}_c)$, which represents the effectiveness of $\mathcal{F}(\mathbf{G}_c)$ for the target output. In details, for the classification task, Distance$(\cdot)$ is the cross-entropy function; and for the regression task, Distance$(\cdot)$ is the root-mean-square error. Sum$(\cdot)$ and Discrete$(\cdot)$ measure the size and discrete degree of $\mathbf{M}_c$, which lead the $\mathbf{M}_c$ to a smaller size.

3. Experiments

3.1. Datasets

In order to verify the validity of the proposed method, the experimental part uses two publicly available datasets of molecules. In addition, to demonstrate that the molecular segmentation method proposed is generalizable to molecular structure data, the datasets were chosen from two fields, chemistry and drug, respectively. The Mutagenicity dataset (Kazius et al., 2005) contains 4337 drug molecule graphs for graph classification, with molecules labeled as having mutagenic and non-mutagenic properties. If the chemical structure of a drug has a mutagenic substructure, it will become a toxic drug (Maron and Ames, 1983). Therefore, the identification of mutagenic structures is an important element of current drug research. This database was constructed using rule derivation and validation criteria for toxic groups with mutagenicity, with the aim of improving the reliability and accuracy of mutagenicity prediction. The Solubility dataset (Delaney, 2004) contains 2287 chemical molecule maps for graph performance regression. The target property is the solubility of the molecule in water. The database was collected using linear regression against nine molecular properties. It is well known that hydrophilic R-OH groups contribute to solubility, and compounds containing more thick rings reduce solubility (Hildebrand, 1916).

Dataset is divided into three components for training, validating, and testing with 8:1:1. The training set and validating set are used to do cross-validation. In our work, 5 folds are set to obtain reliable and stable models. The testing set that never participate in the training of models is used to evaluate trained models performance. Graph structure data are transformed from Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988) using RDKIT (Landrum et al., 2013) package.

3.2. Experimental setup

For each dataset, a SMPNN is trained to predict property, and our explainer is used to explain the predictions. Note that intra-category message passing layers N is 2, parameters θ_u^{l-1} and θ_w^{l-1} in Eq. (3) are built by a fully connected neural network, and one layer message

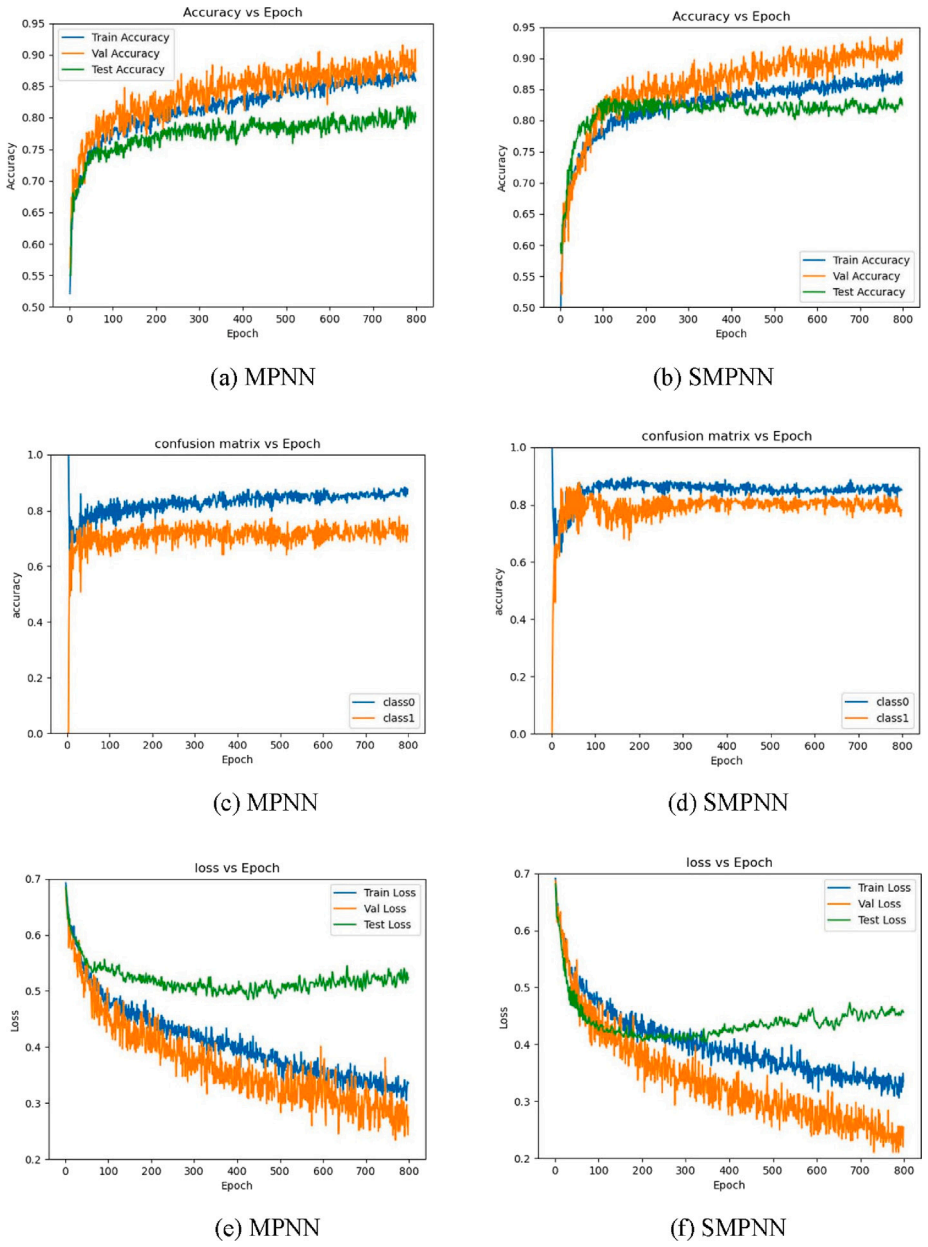

Fig. 4. Mutagenicy dataset peroperty classification prediction accuracy, confusion matrix, and model loss curves with number of iterations based on graph message passing neural network (MPNN) and graph segmentation message passing neural network (SMPNN).

passing between category is applied after intra-category message layers. Global mean pooling is used as the readout operation. A 2-layer fully connected neural network is used to do the classification or regression task at the end of SMPNN. A MPNN with the same settings, which contains number of layers, epochs, and learning rate, is built as comparison. Optuna package is used to optimized the hyperparameters, for example learning rate, in SMPNN and MPNN. Explainer based molecular segmentation is build after SMPNN is trained. The epoch size is set to 100. GNNExplainer is set as the comparison method of explanation. All experiments are conducted using an NVIDIA GetForce GTX 1080Ti with 12 GB memory.

3.3. Results and discussion

3.3.1. Results of molecular segmentation

One result of the molecular segmentation method is shown in Fig. 2 at both node and edge levels. Molecular segmentationing methods are unsupervised and applicable to most molecular structures. More examples of segmentation results from the solubility dataset are shown in Fig. 3. The SMILES sequence of molecules is first used to generate graph topology data, and the molecular structures are classified into different categories by the molecular segmentation method. In the classification graph of nodes, the nodes belonging to the same category are labeled with the same color; in the classification graph of edges, the edges connected by nodes in the same category belong to the same edge category, and the connections between categories are not marked with color, but are equally important. In each molecule, molecules in the same category are labeled with the same color. As can be seen from the figure, the molecular segmentation method proposed in this subsection divides all kinds of molecular structures into different categories in an unsupervised learning mode while preserving the functional group structure as much as possible.

3.3.2. SMPNN performance

SMPNN and MPNN were used in the Mutagenicity dataset for classification prediction of whether a molecule is mutagenic or not, respectively. The analysis of the performance prediction contains the change of the prediction accuracy, and the convergence of the loss with the number of training iterations. The performance prediction results of the two neural networks are shown in Fig. 4. From the overall prediction accuracy, the test set accuracy of the two classes, and the variation curve of the loss value with the number of training iterations in the figure, it can be concluded that: in terms of prediction accuracy, the two neural networks do not differ much in their classification prediction accuracy for this dataset; in terms of network stability, the curve corresponding to the graph message passing network based on the segmentation method is flatter, indicating that its model is more stable in the training process; in terms of network complexity, by counting the number of neighbor node weights trained during the message passing, the number of weight parameters in SMPNN is only 56.8% of that in MPNN. So the graph network based on the segmentation method has fewer parameters in it and lower complexity because of the range limitation of neighbor node selection.

The classification predictions of the completed SMPNN and MPNN on the Mutagenicity dataset after training are represented as confusion matrices, as shown in Fig. 5. The confusion matrix represents the molecular categories predicted by the model horizontally and the actual categories vertically, both of which are of two kinds in this dataset: mutagenicity and non-mutagenicity. From the comparison of the values in the confusion matrix, it can be concluded that SMPNN has higher prediction accuracy compared to MPNN.

The predicted performance regression results of SMPNN and MPNN in the solubility database are shown in Fig. 6. In Fig. 6(a) (b), the loss functions of MPNN and SMPNN decrease with increasing number of iterations, respectively. In terms of numerical analysis, the MPNN model converges to a final loss value of 1.201 and the SMPNN model converges to a final loss value of 0.680. In terms of the convergence rate of the model, the SMPNN converges a little faster than the MPNN. In terms of the stability of the model, there are some raised spikes in the loss curve of MPNN and the curve is not smooth; in contrast, the curve of SMPNN is smoother. In Fig. 6(c)(d), scatter plots of the true and predicted values of the test set are shown, with the horizontal coordinates indicating the true values of molecular solubility and the vertical coordinates indicating the predicted values of the model. The corresponding test set for two models are the same. From the figure, it can be seen that the predicted scatter distribution of SMPNN is closer to the diagonal line than that of MPNN, which indicates that the prediction result of SMPNN is closer to the true value and has better prediction effect. In summary, the SMPNN proposed in this chapter has better convergence performance and better prediction regression effect.

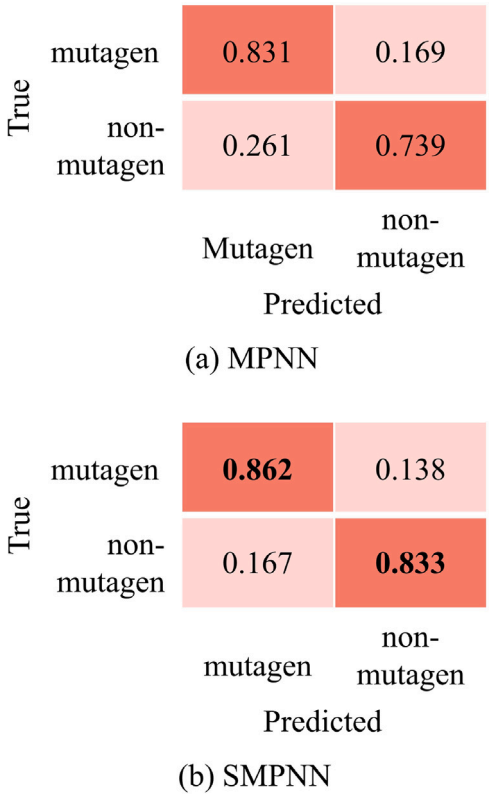

Fig. 5. Mutagenic dataset property classification prediction confusion matrix based on graph message passing neural network (MPNN) and graph segmentation message passing neural network (SMPNN).

Table 1 reports the 100 epochs running time, loss, and Pearson coefficient of MPNN and SMPNN. This comparison is based on the fact that SMPNN and MPNN have the same number of network layers, training data, and learning rate, etc. SMPNN and MPNN contain different numbers of parameters due to the difference in message passing. The running time of SMPNN is 18.5% shorter than that of MPNN. The reason for the shorter running time of SMPNN is SMPNN involves learning within and between molecular segmentation clusters, which reduces the repeated operations required to aggregate neighbor information compared to the indiscriminate learning approach in MPNN where information is aggregated between atoms. Loss and Pearson coefficient are calculated based on the predicted output and true values. The smaller the loss means that the predicted value of the model is closer to the true value, and the closer the Pearson coefficient is to 1, the stronger the correlation between the predicted and true values. SMPNN also has lower loss and higher Pearson coefficient, Which both illustrate that SMPNN achieves more accurate predictions than MPNN. Overall, SMPNN achieves a higher predicted accuracy than MPNN with a lower time cost.

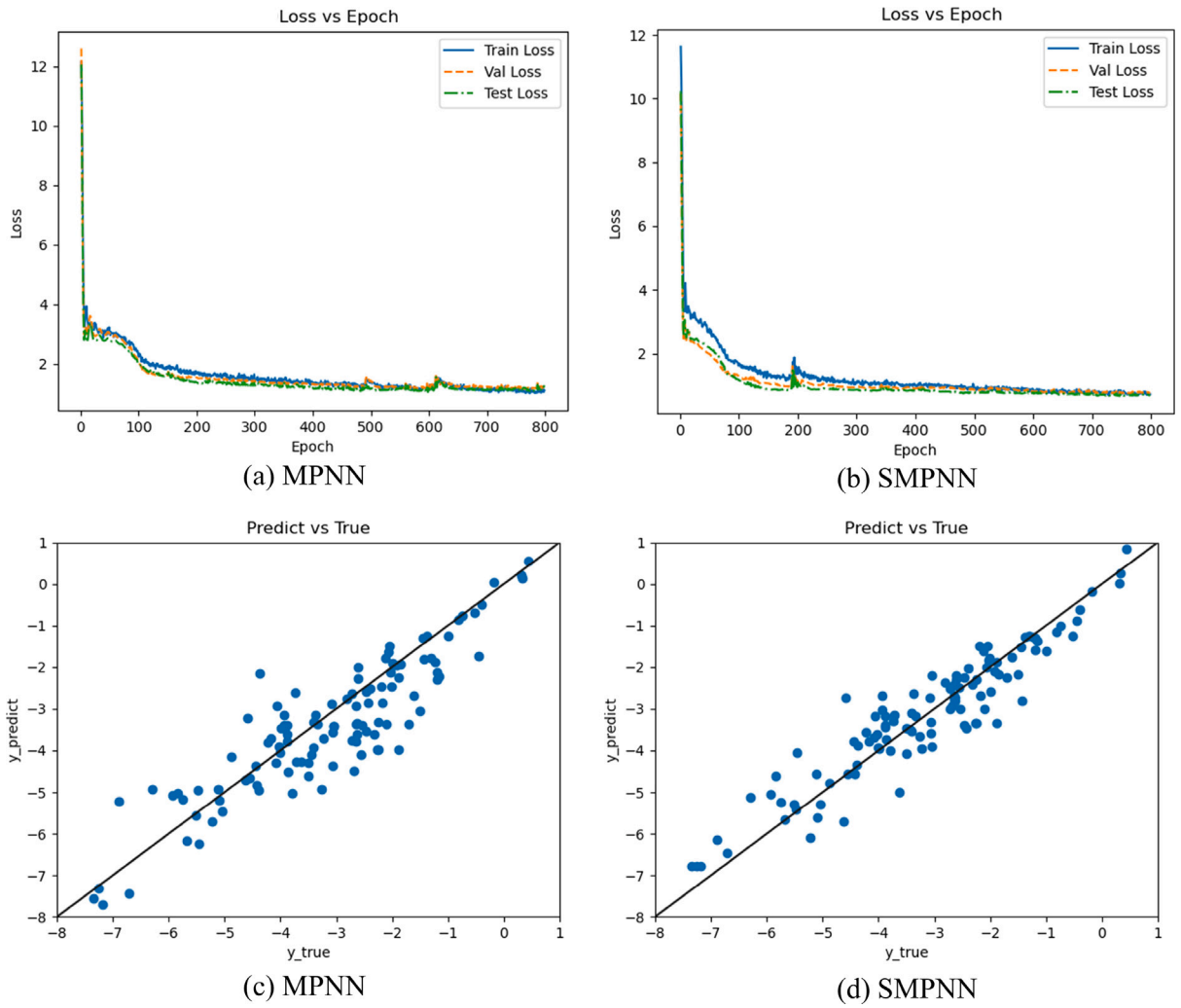

Fig. 6. Solubility dataset property classification prediction accuracy, and model loss curves with number of iterations based on graph message passing neural network (MPNN) and graph segmentation message passing neural network (SMPNN).

Table 1
Running time in 100 epochs, loss, and Pearson coefficient comparisons between graph message passing neural network (MPNN) and graph segmentation message passing neural network (SMPNN).

	MPNN	SMPNN
Time (s)	51.91	**42.29**
Loss	1.156	**0.594**
Pearson	0.904	**0.939**

3.3.3. Explainer performance

The structure-performance relationship explanation of mutagenicity performance in the Mutagenicity dataset is shown in Figs. 7 for the GNNExplainer algorithm and the molecular segmentation based graph explanation algorithm, respectively. Because the Mutagenicity dataset contains atoms N^+ that do not fit the molecular model in RDKIT, the molecular structure of this dataset is plotted in this subsection using the Networkx toolkit, where yellow represents carbon atoms, green represents nitrogen atoms, and blue represents oxygen atoms, using the force-directed layout distribution (Fruchterman-Reingold force-directed) algorithm to generate the node layout. The chemical bonds in a molecule are represented by solid lines between the atoms, and the color and thickness of the lines indicate the degree of influence of the bond on the properties, with the darker and thicker lines representing the more important bonds and the opposite, the thinner and lighter lines representing the less important bonds. For the molecules in the Mutagenicity dataset, molecules with benzene rings andaromatic nitro are more likely to be nonmutagenic (Kazius et al., 2005). Based on this, molecules containing benzene rings and functional groups are partially listed in the dataset to verify and compare the effectiveness of the graph explanation algorithm proposed in this chapter. In Fig. 7(a), the explanation results obtained from the basic weight sequence-based graph explanation algorithm are shown, and in (b), the results obtained from the graph explanation method proposed in this chapter. From the comparison of the figures, it can be seen that the graph explanation algorithm based on molecular segmentation proposed in this chapter can more accurately identify the benzene ring and functional group structures that are important for the performance, and the explained substructures derived from molecular segmentation contain more molecular substructures containing functional groups, for example, the graph explanation method in this chapter can label the entire benzene ring structure, and comparing the graph explainer algorithm with the benzene ring on a bolded bond, the result is clearer and more reasonable from the point of view of the principles and mechanistic knowledge of chemistry.

The structure–property relationships in the Solubility dataset for the two explanation algorithms are shown in Fig. 8. The molecules are plotted by RDKIT, where the importance of the bonds is plotted according to the "bwr" colormap. If the color of the bond is close to blue, it means that the bond has little effect on the performance, and if

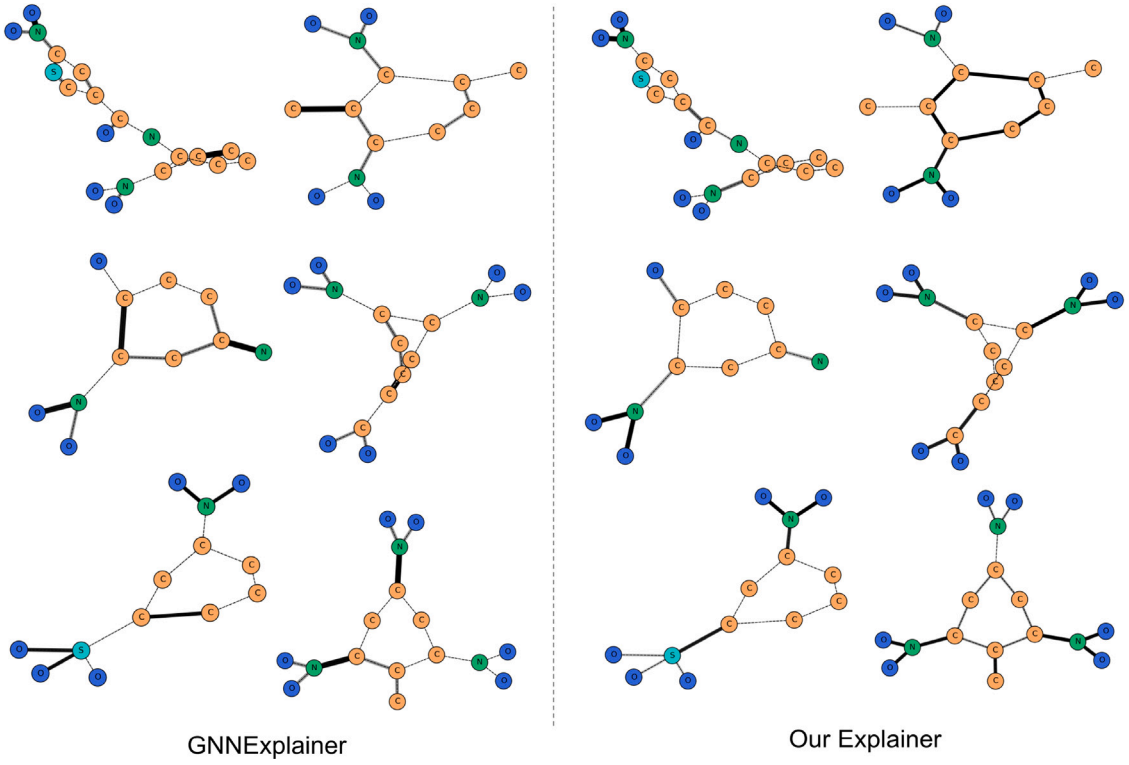

Fig. 7. Diagram of structure–property relationship explanations of graph explain algorithm (left) and graph explain algorithm based on molecular segmentation (right) in Mutagenicity dataset.

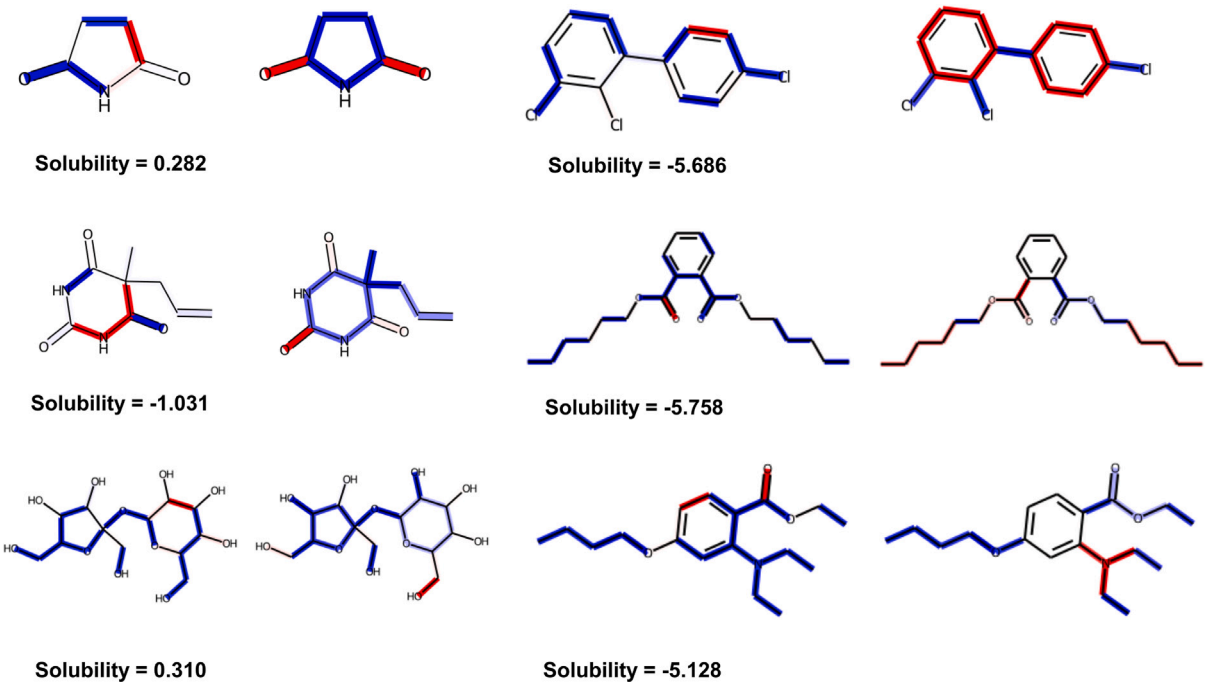

Fig. 8. Diagram of structure–property relationship explanations of graph explain algorithm (left) and graph explain algorithm based on molecular segmentation (right) in Solubility dataset.

it is close to red, it means that the bond has a relatively large effect on the performance. For the molecules listed in the figure, the results of the graph explanation algorithm are shown on the left, and the results of the graph explanation algorithm based on molecular segmentation are shown on the right. Below each group of molecules, the solubility value corresponding to that molecule is marked, and the larger the solubility value, the better the solubility in water. Analyzing the results of the molecular explanation of the graph, the molecule with a solubility of 0.282, using the explanation method proposed in this chapter, it is possible to pinpoint that the benzene ring in the molecule has a negative effect on this solubility and the hydroxyl group has a positive effect on this solubility, because the molecule is more soluble, so the hydrophilicity of the hydroxyl group has a positive effect on the prediction of solubility is in accordance with the mechanistic knowledge; in the explanation of the molecule with a solubility of −5.686, the benzene ring has a positive effect on this solubility and the hydroxyl group has a negative effect, because the solubility of this molecule is low, so the explanation of the result that the important is the benzene ring that does not have hydrophilicity is consistent with the mechanism. The rest of the molecular structure explanations are similar. It can be concluded from several molecular structure explanations that hydroxyl group helps to increase the solubility of the molecule and benzene ring or carbon chain structure has drainage which decreases the solubility of the molecule.

4. Conclusions

In this paper, we construct a GNNs algorithm based on the molecular segmentation method to improve the pattern of message passing in the graph message passing network using the categories delineated by the molecular segmentation method and the substructure explanation mechanism in the graph explanation algorithm, respectively. The segmentation-based graph information network uses first learning the intra-category features containing functional groups, and then learning the overall molecular information. The segmentation-based graph explanation algorithm uses molecular segmentation clusters to define new sequences of weights to represent structure–property relationships. The analysis of the experimental results leads to the following conclusions. The molecular segmentation method, an unsupervised learning method constructed based on the knowledge of molecular structure mechanism, can partition the molecular structure into multiple categories without splitting the functional group structure. The method is simple and effective, and provides a new perspective for molecular structure research. The graph message passing network that combines the molecular segmentation results, using the category distinction after molecular segmentation, improves the order of message passing in it, and not only achieves the same or slightly stronger prediction results than the message passing network of the original learning model, but also reduces the number of parameters that need to be trained in the network. The molecular segmentation results are combined with the graph explanation algorithm to construct a new sequence of weights to learn the importance of the segmented substructure, resulting in an important substructure that is no longer discrete and difficult to explain, and a more direct and accurate explanation of the structure-performance relationship that is more practically meaningful.

CRediT authorship contribution statement

Zhudan Chen: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Dazi Li:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Minghui Liu:** Visualization, Investigation, Validation. **Jun Liu:** Resources, Supervision.

Declaration of competing interest

All authors disclosed no relevant relationships.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62273026.

References

Aldeghi, M., Coley, C.W., 2022. A graph representation of molecular ensembles for polymer property prediction. Chem. Sci. 13 (35), 10486–10498. http://dx.doi.org/10.1039/D2SC02839E.

Alon, U., Yahav, E., 2020. On the bottleneck of graph neural networks and its practical implications. arXiv preprint arXiv:2006.05205.

Amamoto, Y., 2022. Data-driven approaches for structure-property relationships in polymer science for prediction and understanding. Polym. J. 54 (8), 957–967. http://dx.doi.org/10.1038/s41428-022-00648-6.

Bai, Q., Liu, S., Tian, Y., Xu, T., Banegas-Luna, A.J., Pérez-Sánchez, H., Huang, J., Liu, H., Yao, X., 2022. Application advances of deep learning methods for de Novo drug design and molecular dynamics simulation. Wiley Interdiscip. Rev. Comput. Mol. Sci. 12 (3), e1581.

Baldassarre, F., Azizpour, H., 2019. Explainability techniques for graph convolutional networks. arXiv preprint arXiv:1905.13686.

Blondel, J., 2003. Guilds or Functional Groups: Does It Matter? Wiley Online Library.

Cheung, M., Shi, J., Jiang, L., Wright, O., Moura, J.M.F., 2019. Pooling in graph convolutional neural networks. In: Matthews, M.B. (Ed.), Conference Record of the 2019 Fifty-Third Asilomar Conference on Signals, Systems & Computers. pp. 462–466.

Delaney, J.S., 2004. ESOL: Estimating aqueous solubility directly from molecular structure. J. Chem. Inf. Comput. Sci. 44 (3), 1000–1005. http://dx.doi.org/10.1021/ci034243x.

Ertl, P., 2017. An algorithm to identify functional groups in organic molecules. J. Cheminform. 9 (1), 36. http://dx.doi.org/10.1186/s13321-017-0225-z.

Fan, W., Ma, Y., Li, Q., Wang, J., Cai, G., Tang, J., Yin, D., 2022. A graph neural network framework for social recommendations. IEEE Trans. Knowl. Data Eng. 34 (5), 2033–2047. http://dx.doi.org/10.1109/TKDE.2020.3008732.

Fradkin, P., Young, A., Atanackovic, L., Frey, B., Lee, L.J., Wang, B., 2022. A graph neural network approach for molecule carcinogenicity prediction. Bioinformatics 38 (Suppl 1), i84–i91. http://dx.doi.org/10.1093/bioinformatics/btac266.

Fung, V., Zhang, J., Juarez, E., Sumpter, B.G., 2021. Benchmarking graph neural networks for materials chemistry. npj Comput. Mater. 7 (1), 84.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning, Vol. 70. pp. 1263–1272. http://dx.doi.org/10.48550/arXiv.1704.01212.

Hildebrand, J.H., 1916. Solubility. J. Am. Chem. Soc. 38 (8), 1452–1473.

Kazius, J., McGuire, R., Bursi, R., 2005. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48 (1), 312–320. http://dx.doi.org/10.1021/jm040835a.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Landrum, G., et al., 2013. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 8.

Li, Z., Wellawatte, G.P., Chakraborty, M., Gandhi, H.A., Xu, C., White, A.D., 2020. Graph neural network based coarse-grained mapping prediction. Chem. Sci. 11 (35), 9524–9531. http://dx.doi.org/10.1039/D0SC02458A.

Lin, W., Lan, H., Wang, H., Li, B., 2022. Orphicx: A causality-inspired latent variable model for interpreting graph neural networks. http://dx.doi.org/10.48550/arXiv.2203.15209, arXiv arXiv:2203.15209.

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., Zhang, X., 2020. Parameterized explainer for graph neural network. Adv. Neural Inf. Process. Syst. 33, 19620–19631.

Maron, D.M., Ames, B.N., 1983. Revised methods for the salmonella mutagenicity test. Mutat. Res. Environ. Mutagen. Relat. Subj. 113 (3–4), 173–215.

Nt, H., Maehara, T., 2019. Revisiting graph neural networks: All we have is low-pass filters. arXiv preprint arXiv:1905.09550.

Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E., Hoffmann, H., 2018. Discovering molecular functional groups using graph convolutional neural networks. CoRR abs/1812.00265 arXiv:1812.00265.

Rittig, J.G., Ben Hicham, K., Schweidtmann, A.M., Dahmen, M., Mitsos, A., 2023. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput. Chem. Eng. 171, 108153. http://dx.doi.org/10.1016/j.compchemeng.2023.108153.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., Huang, J., 2020. Self-supervised graph transformer on large-scale molecular data. In: 34th Conference on Neural Information Processing System. Vancouver, Canada, http://dx.doi.org/10.48550/arXiv.2007.02835, arXiv arXiv:2007.02835.

Ruiz, L., Gama, F., Ribeiro, A., 2021. Graph neural networks: Architectures, stability, and transferability. Proc. IEEE 109 (5), 660–682. http://dx.doi.org/10.1109/JPROC.2021.3055400.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE Trans. Neural Netw. 20 (1), 61–80.

Schweidtmann, A.M., Rittig, J.G., Weber, J.M., Grohe, M., Dahmen, M., Leonhard, K., Mitsos, A., 2023. Physical pooling functions in graph neural networks for molecular property prediction. Comput. Chem. Eng. 108202. http://dx.doi.org/10.1016/j.compchemeng.2023.108202.

Shearer, J., Castro, J.L., Lawson, A.D.G., MacCoss, M., Taylor, R.D., 2022. Rings in clinical trials and drugs: Present and future. J. Med. Chem. 65 (13), 8699–8712. http://dx.doi.org/10.1021/acs.jmedchem.2c00473.

Shi, X., Hao, K., Chen, L., Wei, B., Liu, X., 2022. Multivariate time series prediction of complex systems based on graph neural networks with location embedding graph structure learning. Adv. Eng. Inform. 54, 101810. http://dx.doi.org/10.1016/j.aei.2022.101810.

Song, Z., Chen, X., Meng, F., Cheng, G., Wang, C., Sun, Z., Yin, W.J., 2020. Machine learning in materials design: Algorithm and application*. Chin. Phys. B 29 (11), 116103. http://dx.doi.org/10.1088/1674-1056/abc0e3.

Tian, R., Li, Y., Wang, X., Li, J., Li, Y., Bei, S., Li, H., 2022. A pharmacoinformatics analysis of artemisinin targets and de Novo design of hits for treating ulcerative colitis. Front. Pharmacol. 13.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

Wardle, D.A., Zackrisson, O., 2005. Effects of species and functional group loss on island ecosystem properties. Nature 435 (7043), 806–810.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28 (1), 31–36.

Wen, H., Su, Y., Wang, Z., Jin, S., Ren, J., Shen, W., Eden, M., 2022. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. AIChE J. 68 (1), e17402. http://dx.doi.org/10.1002/aic.17402.

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., Langer, T., 2020. A compact review of molecular property prediction with graph neural networks. Drug Discov. Today 37, 1–12.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2021. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32 (1), 4–24. http://dx.doi.org/10.1109/TNNLS.2020.2978386.

Yang, Z., Zhong, W., Zhao, L., Chen, C.Y.C., 2022. MGraphDTA: Deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chem. Sci. 13 (3), 816–833. http://dx.doi.org/10.1039/D1SC05180F.

Yasir, A.M., Ma, J., Ouyang, X., Zhao, J., Zhao, Y., Weng, L., Islam, M.S., Chen, Y., Li, Y., 2022. Effects of selected functional groups on nanoplastics transport in saturated media under diethylhexyl phthalate co-contamination conditions. Chemosphere 286, 131965.

Ye, Z., Kumar, Y.J., Sing, G.O., Song, F., Wang, J., 2022. A comprehensive survey of graph neural networks for knowledge graphs. IEEE Access 10, 75729–75741. http://dx.doi.org/10.1109/ACCESS.2022.3191784.

Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. GNN explainer: A tool for post-hoc explanation of graph neural networks. In: In Proceedings of the 33rd Conference on Neural Information Processing Systems. NeurIPS 2019, arXiv:1903.03894.

Yuan, H., Yu, H., Gui, S., Ji, S., 2022. Explainability in graph neural networks: A taxonomic survey. http://dx.doi.org/10.48550/arXiv.2012.15445, arXiv arXiv:2012.15445.

Yuan, H., Yu, H., Wang, J., Li, K., Ji, S., 2021. On explainability of graph neural networks via subgraph explorations. In: International Conference on Machine Learning. PMLR, pp. 12241–12252.

Zhang, J., Wang, Q., Shen, W., 2022a. Message-passing neural network based multi-task deep-learning framework for COSMO-SAC based σ-profile and VCOSMO prediction. Chem. Eng. Sci. 254, 117624. http://dx.doi.org/10.1016/j.ces.2022.117624.

Zhang, J., Wang, Q., Su, Y., Jin, S., Ren, J., Eden, M., Shen, W., 2022b. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChE J. 68 (6), e17634. http://dx.doi.org/10.1002/aic.17634.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. AI Open 1, 57–81.