Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Classification of cold and hot medicinal properties of Chinese herbal medicines based on graph convolutional network

YANG Mengling, LIU Wei*

School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

ARTICLE INFO

Article history
Received 25 August 2024
Accepted 08 October 2024
Available online 25 December 2024

Keywords
Chinese herbal medicine
Cold and hot medicinal properties
Molecular descriptor
Graph convolutional network (GCN)
Medicinal property classification

ABSTRACT

Objective To develop a model based on a graph convolutional network (GCN) to achieve efficient classification of the cold and hot medicinal properties of Chinese herbal medicines (CHMs).

Methods After screening the dataset provided in the published literature, this study included 495 CHMs and their 8 075 compounds. Three molecular descriptors were used to represent the compounds: the molecular access system (MACCS), extended connectivity fingerprint (ECFP), and two-dimensional (2D) molecular descriptors computed by the RDKit open-source toolkit (RDKit_2D). A homogeneous graph with CHMs as nodes was constructed and a classification model for the cold and hot medicinal properties of CHMs was developed based on a GCN using the molecular descriptor information of the compounds as node features. Finally, using accuracy and F1 score to evaluate model performance, the GCN model was experimentally compared with the traditional machine learning approaches, including decision tree (DT), random forest (RF), k-nearest neighbor (KNN), Naïve Bayes classifier (NBC), and support vector machine (SVM). MACCS, ECFP, and RDKit_2D molecular descriptors were also adopted as features for comparison.

Results The experimental results show that the GCN achieved better performance than the traditional machine learning approach when using MACCS as features, with the accuracy and F1 score reaching 0.836 4 and 0.845 3, respectively. The accuracy and F1 score have increased by 0.869 0 and 0.812 0, respectively, compared with the lowest performing feature combination OMER (only the combination of MACCS, ECFP, and RDKit_2D). The accuracy and F1 score of DT, RF, KNN, NBC, and SVM are 0.505 1 and 0.501 8, 0.616 2 and 0.601 5, 0.676 8 and 0.624 3, 0.616 2 and 0.607 1, 0.636 4 and 0.622 5, respectively.

Conclusion In this study, by introducing molecular descriptors as features, it is verified that molecular descriptors and fingerprints play a key role in classifying the cold and hot medicinal properties of CHMs. Meanwhile, excellent classification performance was achieved using the GCN model, providing an important algorithmic basis for the in-depth study of the "structure-property" relationship of CHMs.

*Corresponding author: LIU Wei, E-mail: weiliu@hnucm.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2025.01.008

Citation: YANG ML, LIU W. Classification of cold and hot medicinal properties of Chinese herbal medicines based on graph convolutional network. Digital Chinese Medicine, 2024, 7(4): 356-364.

1 Introduction

Chinese herbal medicines (CHMs), a unique branch of TCM, have been used for over 5 000 years, and play an important role in disease prevention and treatmentare, mainly derived from plant materials [1]. The concept of the four properties of herbs in TCM theory explains the mechanism of action and clinical application of herbs, which provides a theoretical basis for drug compatibility [2].

With the advancements in biochemical and pharmacological research, modern science has begun analyzing the molecular mechanisms of CHMs. For example, through large-scale analyses, researchers have explored the molecular basis of the medicinal properties of herbs [3]. In addition to quantitative methods used to characterize the cold and hot properties of herbs [4], various research methods from the biological sciences were employed to strengthen the scientific foundation of CHMs. In particular, study has used bioinformatics methods to analyze the cold and hot properties of herbs [5]. Meanwhile, metabolomic technology has been applied to the characterization of herbs [6]. The material basis of the cold and hot medicinal properties of CHMs has been further elucidated [7]. However, conducting experiments using biology typically requires a substantial amount of time and resources. The outcomes of these experiments are influenced by various factors, making it challenging to ensure the stability and reproducibility of the results. Based on the above findings, WEI et al. [8] proposed an innovative hypothesis that herbs containing similar substances may have similar cold and hot medicinal properties. To validate this hypothesis and explore the relationship between herbal compounds and their cold and hot properties, researchers have introduced machine learning techniques to the prediction of the cold and hot medicinal properties of CHMs [9-11]. Although machine learning has shown some potential in its initial attempts, it still has limitations, including its high dependence on the quantity and quality of training data and its inability to deal with complex graph data. These limitations restrict its further application and development in classifying medicinal properties.

Deep learning methods have been shown to outperform traditional machine learning methods in predicting the quantum mechanical and physicochemical properties of molecules [12, 13]. These findings have also affected research on the property classification of CHMs, in which artificial intelligence algorithms have become a powerful tool for exploring the complex relationship between the ingredients and properties of CHMs [14-16]. Among them, the graph convolutional network (GCN) [17] has achieved excellent performance in molecular property prediction owing to its advantages in processing data with graph structures [18-22]. A previous study successfully applied the cost-sensitive GCN model to explore the relationship between herbs and their meridians, demonstrating good results [23]. This outcome not only verifies the applicability of GCN in CHMs, but also provides valuable insights for further research. Considering that the chemical composition of herbs is the basis of their efficacy, the classification of CHMs involves analyzing their chemical compositions. Molecular characterization, as a crucial link connecting herbal components to their properties, is essential for predicting the properties of herbs and synthesizing new compounds [24]. Therefore, the compound-based study of the medicinal properties of CHMs using GCN acts as a scientific method for exploring the action mechanism of CHMs in depth. It aims to reveal the chemical components related to the medicinal properties of CHMs and elucidate the targets and trends of their actions on the organism. With sufficient knowledge of the chemical components of CHMs, an urgent need exists for a systematic method to identify the cold and hot properties of CHMs based on their compounds [25]. Exploring the relationship between the medicinal properties and the structural composition of CHMs and constructing a system for characterizing the medicinal components of CHMs can provide a new perspective for theoretical research and scientific interpretation of the medicinal properties of modern CHMs.

This study adopts a novel data representation method that combines herbal compounds and their molecular descriptors to generate node features and uses graph neural networks (GNN) to classify the cold and hot properties of herbs from a chemical perspective, which can more comprehensively capture the chemical properties of herbal compounds.

2 Data and methods

2.1 CHMs dataset

The dataset used in this study was established by WANG et al. [15], which provides a comprehensive and integrated resource, and includes 10 053 compounds from 647 herbs. The therapeutic qualities of CHMs can be broadly classified into four categories: cool, cold, warm, and hot. Another important therapeutic characteristic is neutral. The properties of hot and cold are directly related to the Yin and Yang of the body, where warm and hot belong to Yang and cool and cold to Yin, demonstrating significant impacts on the treatment of diseases. Moreover, warm and cool medicinal properties could be further refined based on the classification of hot and cold [26], aiming to describe the properties of herbs more accurately. In this research, we categorized cold and cool herbs as cold, and hot and warm herbs as hot. The existing information on the properties of herbs was investigated and reorganized to construct a dataset that meets the needs of this study.

Overall, 257 herbs indicated hot properties; 238 herbs, cold properties; 143 herbs, neutral properties, and 8 herbs, unidentified properties. If herbs with neutral and unidentified properties were added to the model training, it would result in a problem that some herbs might not be categorized accurately, thereby affecting the overall model performance. Therefore, 143 herbs with neutral medicinal properties and 8 herbs without defined medicinal properties were removed. All compounds contained in herbs were utilized in this study. The final dataset contained 495 herbs with hot and cold medicinal properties, and 8 075 compounds were labeled for use in the subsequent experiments. Table 1 presents the coding results. If a particular compound is present in a particular herb, the corresponding location is given to a value of 1, if not, a value of 0.

Table 1 Data representation in coded form of herbs and their compounds

Herb name	Property	Embedding					
		1	2	3	4	...	8 075
Yuanzhi (Polygalae Radix)	Hot	0[a]	0	0	0	...	0
Shandougen (Sophorae Tonkinenses Radix et Rhizoma)	Cold	0	0	0	1[b]	...	0
Zhizi (Gardeniae Fructus)	Cold	0	0	0	0	...	1
...	...	...	...	...	...	...	...
Jigucao (Abri Herba)	Cold	1	1	1	1	...	0

[a] represents containing no compound labeled 0. [b] represents containing a certain compound labeled 1.

2.2 Molecular descriptors

2.2.1 Overview of molecular descriptors In chemical research, molecular descriptors are key tools for converting the structural and property information of molecules into quantifiable numerical representations. These descriptors effectively simplify the expression of complex molecular structures and properties, and convert them into computer-processable data formats. Hence, they provide a vital perspective for analyzing the similarities and differences between molecules and predicting their properties. Contemporary research has shown that the properties of CHMs are directly related to their chemical compositions [27], and molecular descriptors serve as a bridge to link the chemical structure to traditional properties. Molecular descriptors are commonly used to extract information regarding CHMs compounds. The structures of compounds characterized by molecular fingerprints or numerical values of descriptors may be directly related to the hot and cold properties of herbs.

2.2.2 Selected descriptors Three distinct molecular fingerprints and descriptors were used to represent drugs in the dataset, which are widely used in studies of CHMs compounds [14, 23]. In particular, the molecular access system (MACCS) [28], extended connectivity fingerprint (ECFP) [29], and two-dimensional (2D) molecular descriptors (RDKit_2D) were computed using the RDKit open-source toolkit [30]. Each of these techniques offers different insights into chemical properties. As a binary fingerprint, the MACCS is comprised of 166 segment definitions. Different molecular substructures are shown in each segment. A 166-bit binary vector is produced by setting a nonexistent location to 0 and an existing substructure location to 1. The ECFP is a descriptor based on a topological molecular fingerprint algorithm, which abstracts the atomic and bond connections inside the molecule as well as the frequency of ring substructures to create a fixed-length binary vector. In this research, we adopted ECFP6, where “6” indicates an atomic environment with a radius of three. A total of 2 048 molecular fingerprints were obtained. The size of the extracted pieces depends on the specified radius; larger radii extract larger fragments, whereas smaller radii extract smaller fragments. The third descriptor, RDKit_2D, is a set of physicochemical descriptors related to molecular structures that are used to quantify the structures and properties of compounds. These descriptors cover a wide range of attributes, such as the size, structure, and electrical characteristics of a molecule. For example, the relative molecular mass, number of hydrogen bond acceptors, and hydrophobicity are included, and these descriptors are commonly employed to analyze and compare molecules.

2.2.3 Feature combinations According to existing research, the selection of atomic representation has a marked impact on the model performance [31]. The selection of the most appropriate feature subset for a specific task is critical. In this study, the three molecular descriptors (MACCS, ECFP6, and RDKit_2D) were used to construct seven distinct descriptor combinations. Two cases were considered: feature combinations without molecular descriptors and feature combinations with only molecular descriptors. These feature combinations were input into the model as node features along with the collated compound data for training (Table 2).

Table 2 Different combinations of molecular descriptors for classifying the cold and hot properties of herbs

Name	Combination of different features
Compound	Compound
OMER	MACCS, ECFP6, RDKit_2D
M	Compound, MACCS
E	Compound, ECFP6
R	Compound, RDKit_2D
ME	Compound, MACCS, ECFP6
MR	Compound, MACCS, RDKit_2D
ER	Compound, ECFP6, RDKit_2D
MER	Compound, MACCS, ECFP6, RDKit_2D

2.3 Graph representation of the data

The graph is defined as $G = (V, E, A)$, where V represents the set of herb nodes, E represents the set of edges between herb nodes, and A is the adjacency matrix used to represent the connection relationship between herb nodes. The a_{ij} denotes the weights and $e_{ij}(e_{ij} \in E)$ denotes the edges between herb nodes. It is assumed to be a binary matrix, which is expressed as follows:

$$A_{ij} = \begin{cases} a_{ij} & \text{If } e_{ij} \in E \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

The degree matrix D of graph G, an $n \times n$ diagonal matrix, v_i $(v_i \in V)$ denotes the herb nodes, is defined as follows:

$$D_{ij} = \begin{cases} \deg(v_i) & \text{If } i = j \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

In addition, a graph has the following properties: (i) similar vector representations occur between adjacent nodes; (ii) vector representations of nodes with comparable characteristics are similar; (iii) vector representations are unaffected by changes in the order of the nodes.

One problem that must be overcome when analyzing CHMs and their compounds is the high complexity, arising from the many-to-many relationship; a single CHM can contain many different compounds, and the same compound can be found in many different herbs. Hence, research on the compounds found in herbs should not be limited to categorizing them according to their degree of hot or cool they are. This simplified classification method overlooks the complex interactions among the compounds of CHMs; therefore, it does not accurately reflect the true medicinal properties of CHMs.

Consequently, this study proposed an approach to understanding the medicinal properties of herbs more precisely. In recent studies, the treatment method for determining the cold and hot properties of compounds based on the cold and hot properties of CHMs faces the issue that the same compound may be classified as different cold and hot properties [14]. It is therefore necessary to maintain the correlation between CHMs and their chemical components for the classification of CHMs. The compound information provides the chemical basis of CHMs, which is conducive to scientifically verifying the rationality and validity of the classification of CHMs. We constructed a graph using herbs as nodes and their chemical molecular information (Figure 1), and calculated molecular descriptors as the characteristics of each herb node. Complex graph-structured data containing information regarding the herbs and compounds were constructed as a consequence. This method can reveal the complexity and medicinal properties of herbal ingredients and provide a more accurate scientific basis for in-depth research into CHMs and their applications.

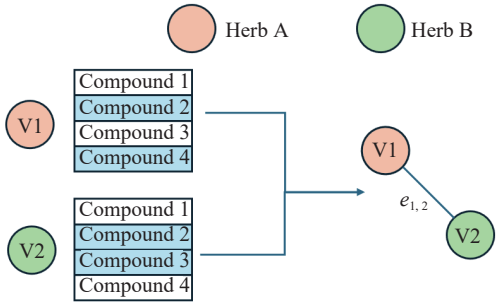

Figure 1 The mapping process of herb nodes and edges
Herb A and Herb B represent two herb nodes. The compounds 1 – 4 represent the compounds contained in Herb A and Herb B, respectively. The blue rectangle in the figure represents that these two herbs have partially identical compounds, and an edge $e_{1,2}$ has been added between V1 and V2.

2.4 GCN

GCN is a deep GNN-based representation learning architecture that defines the convolution operation using a Laplacian matrix [17]. A GCN primarily aims to extract the spatial properties from a topological map, which adopts the properties of the current node and first-order nodes to characterize the new properties of the current node. The prediction results of each node were influenced by nearby nodes according to their association relationships.

The core concept of the GCN algorithm is to spectrally decompose the Laplacian matrix spectrally; that is, to decompose the matrix into a product of the corresponding eigenvalues and eigenvectors. Eigen decomposition is only possible for matrices that can be diagonalized or have multiple linearly uncorrelated eigenvectors. In graph representation learning, the Laplacian matrix is a symmetric matrix defined as $L = D - A$, which is expressed as follows:

$$L_{ij} = \begin{cases} \deg(v_i) & \text{If } i = j \\ -1 & \text{If } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

where D is a degree matrix representing the degree of each herb node. Specifically, D is a diagonal matrix whose diagonal elements d_i represent the degrees of herb node i, A is an adjacency matrix that represents the connection between herb nodes, and L is the Laplacian

matrix that represents the relationship between the herb nodes. The Laplacian matrix can be used to transform the features as follows:

$$L' = 2L/\lambda_{\max} - I \tag{4}$$

where $\lambda_{\max}$ is the maximum eigenvalue of the Laplacian matrix and I the identity matrix. This transformation maps the eigenvalues of the Laplacian matrix to the interval [–1, 1] to make the calculation more stable. The transformed Laplacian matrix is normalized as follows:

$$L^{sym} = D^{-\frac{1}{2}} L' D^{-\frac{1}{2}} \tag{5}$$

where $D^{-\frac{1}{2}}$ denotes the inverse square root of matrix D.

The GCN convolution operation is expressed as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{6}$$

Here, $H^{(l)}$ denotes the herb node feature matrix of layer l, σ the activation function, $W^{(l)}$ the weight matrix of layer l, $\tilde{A} = A + I_n$ the adjacency matrix summed with the matrix of self-loops, and $\tilde{D}$ the degree matrix whose diagonal elements are $\tilde{A}$. In the GCN convolution operation, $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ normalizes the adjacency matrix such that the eigenvalue of each herb node is affected by its neighboring nodes, and $H^{(l)}W^{(l)}$ represents the linear transformation of herb node features to obtain a new feature representation. By repeatedly stacking the GCN convolutional layers, a deep GNN model can be obtained for performing various tasks on the graph data.

The final GCN model is expressed as follows:

$$Z = f(X,A) = \text{softmax}\left(\hat{A}ReLU\left(\hat{A}XW^{(0)}\right)W^{(1)}\right) \tag{7}$$

where Z is the output result, $\hat{A} = A + I$ the updated adjacency matrix, and I the identity matrix. $W^{(0)}$ and $W^{(1)}$ are learnable weight matrices, and the input herb node feature matrix X was multiplied by the weight matrix $W^{(0)}$ to obtain the intermediate feature representation. The feature representation was multiplied by $\hat{A}$ to perform a $ReLU$ activation function operation, following which the results were multiplied by the weight matrix $W^{(1)}$ to obtain the final feature representation. The feature representation was then normalized using softmax to obtain the final CHMs classification result. The GCN structure employed in this study is shown in Figure 2. And the workflow of the GCN was used to classify the cold and hot properties of herbs (Figure 3).

2.5 Evaluation indices

The accuracy and F1 score were used to evaluate the performance of the model in classification tasks. This study evaluates the performance of the proposed GCN model with five traditional machine learning algorithms in the task of cold and hot classification of herbs using accuracy and F1 score. These five traditional machine learning

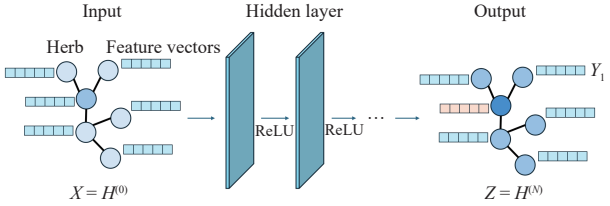

Figure 2 GCN network structure

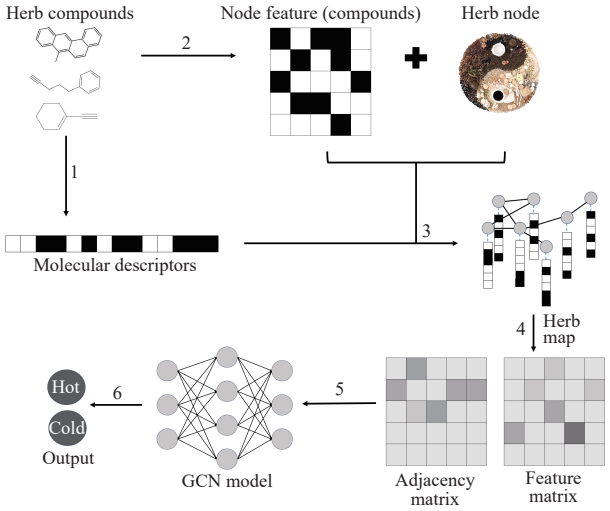

Figure 3 Flowchart of GCN-based classification of the cold and hot properties of herbs

1, calculation of the molecular descriptors of the compounds. 2, conversion of compounds into vectors. 3, formation of herb nodes with molecular descriptors and edges based on shared compounds. 4, collection of adjacency and feature matrices. 5, input of matrices into the GCN model. 6, output results of the model.

algorithms are decision tree (DT), random forest (RF), k-nearest neighbor (KNN), Naïve Bayes classifier (NBC), and support vector machine (SVM). The true categories of the samples are cold and hot. Here, the positive class which consists of herbs was classified as having cold properties, and the negative class which consists of herbs as having hot properties. The number of true positives (TP) represents the number of herbs with cold properties predicted by the model with cold properties. The number of false negatives (FN) represents the number of herbs with cold properties predicted by the model with having hot properties. The number of false positives (FP) represents the number of herbs with hot properties predicted by the model with cold properties. Finally, the number of true negatives (TN) represents the number of herbs with hot properties predicted by the model with hot properties.

Accuracy is the proportion of correctly classified samples relative to the total number of samples and is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{8}$$

The F1 score combined the precision and recall of the model and was used to measure the balanced

performance for positive and negative samples. The precision evaluated the number of positive samples predicted by the model as actually positive, indicating the accuracy of the model's prediction of positive examples. Recall assessed the proportion of samples correctly predicted as positive by the model among all true positive samples as well as the predicted positive samples of the model. These metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

3 Results

3.1 Impacts of molecular descriptors

We compared the performances of different combinations of molecular descriptors as features using cross-validation with different folds. The experimental results indicated that the best results were achieved with 10-fold cross-validation (Table 3). When using combination M (Compound, MACCS) as model input, the accuracy and F1 score were maximized, achieving values of 0.836 4 ± 0.020 6 and 0.845 3 ± 0.021 9, respectively. The performance generated when using the combination OMER (only the combination of MACCS, ECFP, and RDKit_2D) to represent features was poor, yielding accuracy and F1 scores of 0.749 5 ± 0.026 3 and 0.764 1 ± 0.025 8, respectively.

Table 3 Accuracy and F1 score of different feature combinations

Name	Accuracy			F1 score		
	3-fold	5-fold	10-fold	3-fold	5-fold	10-fold
Compound	0.781 1 ± 0.019 0	0.804 0 ± 0.021 8	0.828 3 ± 0.011 1	0.788 4 ± 0.016 6	0.816 0 ± 0.019 3	0.837 7 ± 0.011 4
OMER	0.710 4 ± 0.041 5	0.737 4 ± 0.032 6	0.749 5 ± 0.026 3	0.723 3 ± 0.032 5	0.752 7 ± 0.030 9	0.764 1 ± 0.025 8
M	0.771 0 ± 0.026 5	0.798 0 ± 0.019 2	0.836 4 ± 0.020 6	0.777 6 ± 0.027 3	0.808 5 ± 0.017 5	0.845 3 ± 0.021 9
E	0.764 3 ± 0.004 8	0.765 7 ± 0.039 1	0.816 2 ± 0.017 4	0.772 4 ± 0.010 2	0.777 0 ± 0.036 9	0.825 9 ± 0.019 7
R	0.771 0 ± 0.045 4	0.806 1 ± 0.034 0	0.828 3 ± 0.024 7	0.783 7 ± 0.040 6	0.819 2 ± 0.028 2	0.839 5 ± 0.025 7
ME	0.750 8 ± 0.009 5	0.796 0 ± 0.022 5	0.822 2 ± 0.015 8	0.758 2 ± 0.006 1	0.809 6 ± 0.022 1	0.835 2 ± 0.017 7
MR	0.764 3 ± 0.045 4	0.802 0 ± 0.018 7	0.833 3 ± 0.032 3	0.775 9 ± 0.041 8	0.812 8 ± 0.018 8	0.842 6 ± 0.034 3
ER	0.750 8 ± 0.031 2	0.804 0 ± 0.029 7	0.807 1 ± 0.026 9	0.758 8 ± 0.036 7	0.814 2 ± 0.030 0	0.821 2 ± 0.025 4
MER	0.760 9 ± 0.039 0	0.775 8 ± 0.022 5	0.799 0 ± 0.038 5	0.772 1 ± 0.034 7	0.789 4 ± 0.018 2	0.810 8 ± 0.037 0

3.2 Comparison of different models

This section presents a comparison of the proposed GCN model with five traditional machine-learning algorithms: DT, RF, KNN, NBC, and SVM (Figure 4). The experimental results showed that the performance metrics of the GCN method were significantly superior to those of the traditional machine learning algorithms. In most feature combinations, RF and SVM indicated the second-best performance, while DT performs poorly. The GCN model has end-to-end learning capabilities and can learn node representations and classification models directly from the original node features and adjacency matrices. In contrast, traditional machine learning algorithms often require manual feature extraction, which may generate human bias and limit the model performance. Furthermore, these traditional algorithms often have difficulty processing graph-structured data due to the focus on processing traditional tabular data and the failure to capture complex relationships in the graph data.

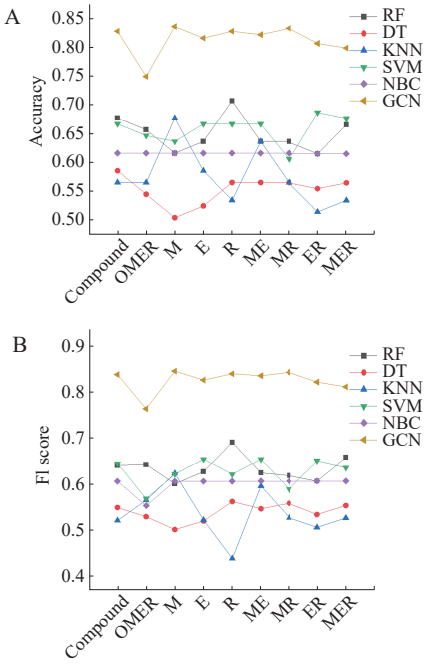

Figure 4 Comparison of the accuracy and F1 score with different methods

A, accuracy. B, F1 score.

4 Discussion

4.1 Impact of molecular descriptors

CHMs are complex drugs that contain many active molecules with highly diverse chemical structures. As a result, selecting appropriate molecular descriptors is crucial for studying CHMs. Different molecular descriptors can extract the characteristics of molecules from different perspectives, including but not limited to, atomic information, structural topology, and charge distribution [32]. Integration of these diverse molecular descriptors into the node characteristics of a molecular graph constituted by herbs can help describe the structures and properties of herbal molecules more comprehensively.

Here, we explored the performance of each descriptor in the classification models in depth by evaluating the selection of three different molecular descriptors. The experimental results indicated that the MACCS fingerprint outperformed the ECFP6 when used to describe the compounds in CHMs. These findings emphasize the crucial importance of compound structural information in classifying the cold and hot properties of herbs. As the ECFP6 contains a 2 048-bit vector, much larger than the 166-bit MACCS fingerprint, it may introduce features that are irrelevant to the classification of cold and hot properties, thereby interfering with the recognition and learning of key features of the model and possibly decreasing the accuracy. Additionally, the RDKit_2D descriptor performed better than the ECFP6, as it contains multiple key features and properties, such as the molecular mass, charge, aromaticity, number of alicyclic rings, and relative molecular mass, which are important factors for classifying the cold and hot properties of herbs [7]. Molecular descriptors were used to convert the structural information of the chemical components of herbs into numerical features, which can facilitate the rapid and accurate extraction of key molecular characteristics, thereby enhancing the effectiveness of reflecting the chemical composition of the herbs.

Therefore, comparing the prediction accuracy, stability, and interpretability of different molecular descriptors is key to optimizing the model performance. Such comparisons are also of great significance for understanding the activity and functions of herb molecules. Further research on the impact of the selection of molecular descriptors on the classification model results will promote the model performance and provide key guidance for the optimal design of herbal molecules and the discovery of new drugs.

4.2 Advantages of GCN in the classification

Compared with traditional machine learning methods, the GCN showed significant advantages in processing graph data, serving as an approach for identifying potential common patterns of cold and hot properties within big datasets of compounds [33]. In particular, the GCN can effectively handle the many-to-many relationships among herbs and important compounds in the analyses of herbal compounds, because the constructed herb representation graph is a complex structure comprising many herb nodes and the interaction edges between them. In this structure, each node represents an herb, and the edges demonstrate the co-occurrence of compounds in different herbs. Traditional machine-learning methods often fail to directly process graph-structured data. The GCN continuously updates the representation of each node by iteratively aggregating the features of neighboring nodes and by fusing local neighborhood information and node features in the graph. This aggregation process enables each node to obtain information from surrounding nodes and effectively integrate this information into its own representation. Thus, a GCN can accurately capture the interactions and relationships among herbs and compounds.

4.3 Future works and challenges

Integrating different molecular descriptors into the node features of the herb representation graph can significantly improve the classification model results, increasing the accuracy and efficiency of research on herbal molecules and providing essential support for the application and development of herbal molecules. In the future, this classification model could be applied to other key topics in the field of CHMs, such as the classification of herb efficacy, thereby expanding its application scope. In summary, the use of GNN techniques to classify the cold and hot properties of herbs has far-reaching research implications and potential applications in the field of CHMs. As the GNN technology continues to mature and gain popularity in the field of artificial intelligence, its potential applications in the field of CHMs will expand. Continuous improvement and expansion of this method are anticipated to provide more precise and comprehensive support for the research and application of herbs. This is expected to promote further development and innovation in the field of TCM. Unfortunately, not all constituent compounds of a herb could be fully recognized. Moreover, each compound in herbal medicine is typically used individually in experiments, but interactions may occur and the main contributions between constituent compounds are not well understood. Given that the present study only considered the use of compounds in determining the connectivity between CHMs and in analyzing the medicinal properties of CHMs from the perspective of compounds, it cannot be separated from the compounds themselves. Looking ahead, the addition of compound nodes and the enhancement of attention mechanism

learning within the model could be considered. This would enable the model to identify nodes with a greater impact on the medicinal properties of CHMs among the many compound nodes and adjust the weight of the connection between the nodes accordingly.

5 Conclusion

An effective classification model was successfully developed by training and validation on the herb compound dataset in this study, which has significant effects on studies and practical applications of the cold and hot properties of herbs. These results will promote the understanding and expansion of the mechanism of CHMs. Many studies of disease mechanisms have focused on the effects of CHMs on cell proliferation and toxicity. By verifying the feasibility of classifying the medicinal properties of CHMs based on compounds, the nature and structure of CHMs compounds can be analyzed. These findings allow for understanding and exploring the action of the CHMs mechanism.

Fundings

Hunan Provincial Natural Science Foundation (2022JJ30438), Natural Science Foundation of Changsha (kq2202260), and Hunan Province Traditional Chinese Medicine Research Project (B2023039).

Competing interests

The authors declare no conflict of interest.

References

[1] ZHANG SQ, JIANG XX, LI JC. Traditional Chinese medicine in human diseases treatment: new insights of their potential mechanisms. Anatomical Record, 2023, 306(12): 2920–2926.

[2] WANG M, SUN YP, WANG ZB, et al. Comment and prospect of research on TCM nature and flavour theory. China Journal of Traditional Chinese Medicine and Pharmacy, 2021, 36(2): 625–628.

[3] YUAN ZZ, PAN YY, LENG T, et al. Progress and prospects of research ideas and methods in the network pharmacology of traditional Chinese medicine. Journal of Pharmacy & Pharmaceutical Sciences, 2022, 25: 218–226.

[4] WANG XR, CAO TT, TIAN XM, et al. Quantification of "cold-hot" medicinal properties of Chinese medicines based on primary metabolites and fisher's analysis. Computational and Mathematical Methods in Medicine, 2022, 2022: 5790893.

[5] WANG Y, ZHOU L, ZHANG SJ, et al. Identification of pathways and genes associated with cold and hot properties of Chinese materia Medica based on bioinformatics analysis. Phytochemistry Letters, 2020, 38: 70–77.

[6] WANG YY, SUN YP, YANG BY, et al. Application of metabolomics and network analysis to reveal the ameliorating effect of four typical "hot" property herbs on hypothyroidism rats. Frontiers in Pharmacology, 2022, 13: 955905.

[7] FU XJ, MERVIN LH, LI XB, et al. Toward understanding the cold, hot, and neutral nature of Chinese medicines using in silico mode-of-action analysis. Journal of Chemical Information and Modeling, 2017, 57(3): 468–483.

[8] WEI GH, FU XJ, HE XY, et al. Cold-hot nature identification based on GC similarity analysis of Chinese herbal medicine ingredients. RSC Advances, 2021, 11(42): 26008–26015.

[9] WEI GH, FU XJ, WANG ZG. Similarity measurement of Chinese medicine ingredients for cold-hot nature identification. TMR Modern Herbal Medicine, 2019, 2(4): 183.

[10] WEI GH, JIA RH, KONG ZY, et al. Cold-hot nature identification of Chinese herbal medicines based on the similarity of HPLC fingerprints. Frontiers in Chemistry, 2022, 10: 1002062.

[11] WEI GH, QIU M, WANG ZG. Multi-wavelength HPLC fingerprint similarity metric for cold-hot nature identification of Chinese herbal medicines. Arabian Journal of Chemistry, 2023, 16(5): 104690.

[12] FENG HJ, QIN LL, ZHANG BX, et al. Prediction and interpretability of melting points of ionic liquids using graph neural networks. ACS Omega, 2024, 9(14): 16016–16025.

[13] STUYVER T, COLEY CW. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. The Journal of Chemical Physics, 2022, 156(8): 1–13.

[14] WANG ML, LI L, YU CY, et al. Classification of mixtures of Chinese herbal medicines based on a self-organizing map (SOM). Molecular Informatics, 2016, 35(3/4): 109–115.

[15] WANG YY, JAFARI M, TANG Y, et al. Predicting meridian in Chinese traditional medicine using machine learning approaches. PLoS Computational Biology, 2019, 15(11): e1007249.

[16] WEI GH, FU XJ, WANG ZG. Nature identification of Chinese herbal medicine compounds based on molecular descriptors. Journal of AOAC International, 2021, 104(6): 1754–1759.

[17] KIPF TN, WELLING M. Semi-supervised classification with graph convolutional networks. arXiv, 2016. Available from: http://arxiv.org/abs/1609.02907v4.

[18] HUNG C, GINI G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Molecular Diversity, 2021, 25(3): 1283–1299.

[19] LEE SM, LEE M, GYAK KW, et al. Novel solubility prediction models: molecular fingerprints and physicochemical features vs graph convolutional neural networks. ACS Omega, 2022, 7(14): 12268–12277.

[20] WEBER JK, MORRONE JA, BAGCHI S, et al. Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. Journal of Computer-Aided Molecular Design, 2022, 36(5): 391–404.

[21] HUANG MT, LOU CF, WU ZR, et al. In silico prediction of UGT-mediated metabolism in drug-like molecules via graph neural network. Journal of Cheminformatics, 2022, 14(1): 46.

[22] LIU HC, PENG W, DAI W, et al. Improving anti-cancer drug response prediction using multi-task learning on graph convolutional networks. Methods, 2024, 222: 41–50.

[23] YEH HY, CHAO CT, LAI YP, et al. Predicting the associations between meridians and Chinese traditional medicine using a cost-sensitive graph convolutional neural network. International Journal of Environmental Research and Public Health,

2020, 17(3): 740.
[24] CHUANG KV, GUNSALUS LM, KEISER MJ. Learning molecular representations for medicinal chemistry. Journal of Medicinal Chemistry, 2020, 63(16): 8705–8722.
[25] ZHANG WJ, HUAI Y, MIAO ZP, et al. Systems pharmacology for investigation of the mechanisms of action of traditional Chinese medicine in drug discovery. Frontiers in Pharmacology, 2019, 10: 743.
[26] GUO J, WANG JX, IINO K, et al. Quantitative and molecular similarity analyses of the metabolites of cold- and hot-natured Chinese herbs. Evidence-Based Complementary and Alternative Medicine, 2021, 2021: 6646507.
[27] XU ZY, DONG M, YIN SP, et al. Why traditional herbal medicine promotes wound healing: research from immune response, wound microbiome to controlled delivery. Advanced Drug Delivery Reviews, 2023, 195: 114764.
[28] DURANT JL, LELAND BA, HENRY DR, et al. Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences, 2002, 42(6): 1273–1280.
[29] ROGERS D, HAHN M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 2010, 50(5): 742–754.
[30] rdkit.org [Internet]. RDKit: Open-source cheminformatics. Available from: https://www.rdkit.org/.
[31] WOJTUCH A, DANEL T, PODLEWSKA S, et al. Extended study on atomic featurization in graph neural networks for molecular property prediction. Journal of Cheminformatics, 2023, 15(1): 81.
[32] FERNÁNDEZ-TORRAS A, COMAJUNCOSA-CREUS A, DURAN-FRIGOLA M, et al. Connecting chemistry and biology through molecular descriptors. Current Opinion in Chemical Biology, 2022, 66: 102090.
[33] ZHAO QC, YANG MY, CHENG ZJ, et al. Biomedical data and deep learning computational models for predicting compound-protein relations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(4): 2092–2110.

基于图卷积网络的中药寒热属性分类研究

杨梦玲, 刘伟*

湖南中医药大学信息科学与工程学院, 湖南 长沙 410208, 中国

【摘要】目的 为了对中药寒热属性进行高效分类，提出了基于图卷积网络（GCN）的分类模型。**方法** 本研究在对已发表文献提供的数据集进行筛选后，最后纳入了 495 种中药及其 8 075 个化合物数据。使用三种分子描述符来表示化合物，分别是分子访问系统（MACCS）、扩展连通性指纹（ECFP）和 RDKit 开源工具包计算的二维（2D）分子描述符（RDKit_2D），构建以中药为节点的同质图，并以化合物分子描述符信息为节点特征，基于图卷积网络提出一种中药寒热属性分类模型。最后，采用准确率和 F1 值评估模型性能，将 GCN 模型与决策树（DT）、随机森林（RF）、K-邻近（KNN）、朴素贝叶斯（NBC）和支持向量机（SVM）进行对比实验，并将 MACCS、ECFP 和 RDKit_2D 分子描述符作为特征进行对比实验。**结果** 实验结果表明，相较于机器学习方法，GCN 取得了较好的性能，使用 MACCS 作为特征准确率和 F1 值分别达到了 0.836 4 和 0.845 3，并且与性能最低的特征组合 OMER（仅是 MACCS、ECFP、RDKit_2D 的组合）相比，准确率和 F1 值分别提升了 0.869 0 和 0.812 0。而 DT、RF、KNN、NBC 和 SVM 的准确率和 F1 值分别为 0.505 1 和 0.501 8、0.616 2 和 0.601 5、0.676 8 和 0.624 3、0.616 2 和 0.607 1、0.636 4 和 0.622 5。**结论** 本研究通过引入分子描述符作为特征，验证了在对中药寒热属性进行分类时，分子描述符与指纹起到了关键作用。同时，利用 GCN 模型实现了出色的分类性能，为深入研究中药的“结构-性质”关系提供了重要的算法依据。

【关键词】中药；寒热药性；分子描述符；图卷积网络；药性分类