


 Cite this: *RSC Adv.*, 2021, **11**, 26008

# Cold–hot nature identification based on GC similarity analysis of Chinese herbal medicine ingredients

 Guohui Wei, <sup>ab</sup> Xianjun Fu,<sup>a</sup> Xueying He,<sup>b</sup> Peng Qiu,<sup>b</sup> Lu Yue,<sup>b</sup> Rong Rong<sup>\*a</sup> and Zhenguo Wang<sup>\*a</sup>

The theory of cold–hot nature of Chinese herbal medicines (CHMs) is the core theory of CHM. It has been found that the volatile oil ingredients in CHMs are closely related to their cold–hot nature. Guided by the scientific hypothesis that “CHMs with similar component substances should have similar medicinal natures”, exploration of the intelligent identification of the cold–hot nature of CHMs based on the similarity of their volatile oil ingredients has become a research focus. Gas chromatography (GC) chemical fingerprints have been widely used in the separation of volatile oil ingredients to analyze the cold–hot nature of CHMs. To verify the above hypothesis, in this work, we study the quantification of the similarity of the volatile oil ingredients of CHMs to their fingerprint similarity and explore the relationship between the volatile oil ingredients of CHMs and their cold–hot nature. In this study, we utilize GC technology to analyze the chemical ingredients of 61 CHMs that have a clear cold–hot nature (including 30 ‘cold’ CHMs and 31 ‘hot’ CHMs). Using the constructed fingerprint dataset of CHMs, a distance metric learning algorithm is applied to measure the similarity of the GC fingerprints. Furthermore, an improved *k*-nearest neighbor (kNN) algorithm is proposed to build a predictive identification model to identify the cold–hot nature of CHMs. The experimental results prove our inference that CHMs with similar component substances should have similar medicinal natures. Compared with existing classical models, the proposed identification scheme has better predictive performance. The proposed prediction model is proved to be effective and feasible.

Received 30th May 2021

Accepted 14th July 2021

DOI: 10.1039/d1ra04189d

[rsc.li/rsc-advances](http://rsc.li/rsc-advances)

## 1. Introduction

As one of the core elements of traditional Chinese medicine (TCM), the nature theory of Chinese herbal medicines (CHMs) has attracted the attention of scholars and research institutions for many years. The nature of CHMs can be classified into four types, *i.e.*, cool, cold, hot, and warm, of which the cold or hot nature is an important part of TCM nature theory.<sup>1,2</sup> The principle of “treating a hot syndrome with a cold-nature medicine and treating a cold syndrome with a hot-nature medicine” indicates that cold–hot medicinal nature theory is an important basis for TCM treatment in regulating the balance between Yin and Yang in the human body, and the application of the cold–hot medicine nature theory is effective for clinical treatment using TCM.<sup>3</sup>

Nature studies from different perspectives have been proposed to reveal the scientific significance of CHM nature.

Some current investigations have focused on analyzing the cold–hot nature of CHMs based on their component substances. Chemical component elements were included to construct a three-element mathematical analysis model for difference analysis for the biological characterization of cold–hot nature.<sup>4</sup> Some research has analyzed the cold–hot nature of CHMs using animal behavioral methods. One study found that cold-nature medicine can regulate the body temperature of rats with yeast-induced fever.<sup>5</sup> Another study found that cold–hot nature was closely related to energy metabolism rates.<sup>6</sup> Some research has analyzed the cold–hot nature of CHMs using bio-informatics methods.<sup>7,8</sup> Liang *et al.* used a molecular network to analyze the cold–hot nature of CHMs.<sup>7</sup> They found that inflammation and immunity regulation were more related to CHMs with the hot nature, and cold-nature CHMs possessed the tendency to impact cell growth, proliferation and development. Our group concluded that compounds associated with a cold nature had a sedative function, associated with “mental and behavioral disorder” diseases, while compounds associated with a hot nature had cardio-protection function, associated with “endocrine, nutritional and metabolic diseases” and “diseases of the circulatory system”.<sup>8</sup>

<sup>a</sup>Key Laboratory of Theory of TCM, Ministry of Education of China, Shandong University of Traditional Chinese Medicine, Jinan 250355, China

<sup>b</sup>College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan 250355, China. E-mail: [rosierong@163.com](mailto:rosierong@163.com); [zhenguo@126.com](mailto:zhenguo@126.com)



The bioactivity of a CHM<sup>1</sup> is determined by its composition, and the bioactivities of CHMs are the key to identifying their medicinal nature. Thus, the material composition of CHMs indirectly determines their nature. Current investigations of medicinal nature are focused on revealing the connection between the nature of a CHM and its material composition. Studies of the chemical basis of CHM nature have shown that hot CHMs contained volatile oil components, while cold CHMs contained glycosides.<sup>9</sup> The medicinal nature of CHMs is closely related to their chemical components. For example, CHMs containing aromatic components in their volatile oil often have a hot nature.<sup>10</sup> Chemical fingerprinting techniques have been used to analyze the chemical components of CHMs.<sup>11,12</sup> Therefore, researchers have explored identification of the cold-hot nature of CHMs using their chemical fingerprints.<sup>13</sup>

Generally, discrimination of the cold-hot nature of CHMs consists of two parts: feature representation and nature classification. Feature representation uses the original effects of the CHMs, metabolomics methods, molecular descriptors or fingerprint technology to extract the characteristics of the CHMs. Nature classification requires the use of classical machine learning classifiers or constructed classifiers to discriminate the cold-hot nature of CHMs. The original effects of CHMs are an effective characteristic expression. Xue's research group explored the original efficacy features of CHMs and used classical classifiers (such as an artificial neural network) to classify the nature of unknown CHMs.<sup>11,12</sup> Metabolomics methods are also applied to represent CHMs. Nie *et al.* studied the metabolomics features of CHMs and constructed a random forest model to discriminate the nature of unknown CHMs.<sup>14</sup> Molecular descriptor technology is an important method for analyzing cold-hot nature. Long *et al.* analyzed the molecular descriptors of compounds of 284 CHMs with clear medicinal natures, and explored a combination system for predicting the cold-hot nature of other CHMs.<sup>15</sup> Other methods, such as proton nuclear magnetic resonance spectroscopy (<sup>1</sup>H-NMR), are used to investigate the features of CHMs. Li *et al.* studied the characteristics of CHMs using <sup>1</sup>H-NMR and applied pattern recognition techniques to analyze the unknown nature of CHMs.<sup>16</sup> Chemical fingerprints can be used to characterize the overall composition of CHMs. Our group studied multi-solvent ultraviolet fingerprints for cold-hot nature identification.<sup>13</sup>

As mentioned above, the discrimination of the cold-hot nature of CHMs has been studied extensively. However, chemical fingerprint technology has not been studied in depth. Our previous research focused on the cold-hot nature discrimination of CHMs based on UV spectra.<sup>13</sup> However, according to the existing research results, studies on the chemical basis of CHM nature have shown that hot CHMs contain volatile oil components, while cold CHMs contain glycosides.<sup>9,10</sup> The volatile oil information of CHMs can be extracted using gas chromatography. It is possible to obtain a better discrimination rate for cold-hot nature using gas chromatography of CHMs. Furthermore, most investigations used existing classical algorithms to construct prediction models, which would result in poor classification. Designing a classifier based on the characteristics of

the fingerprint data may improve the discrimination performance. In this work, gas chromatography technology is applied to extract the characteristic information of CHM ingredients. Using the chemical fingerprint data of CHMs from our research group, a distance metric learning algorithm is constructed to measure the similarity of gas fingerprints, and a prediction model is built to identify the cold-hot nature of CHMs.

## 2. Materials and methods

### 2.1 TCM dataset

61 representative CHMs are analyzed in this study, of which 30 CHMs are 'cold' medicines and the others are 'hot' medicines.<sup>13</sup> All CHMs have been marked in the classical texts *Chinese Materia Medica* and *Shen Nong's Herbal Classic*. GC technology with the steam distillation method was used to obtain the characteristic information of the CHM ingredients. The instrument used in the experiments was a GC6890N GC with a chromatographic data processing system; the reagents were distilled water, anhydrous Na<sub>2</sub>SO<sub>4</sub>, and ethyl acetate. Our group recorded the fingerprint information of a total of 61 CHMs for identification of their nature. Fig. 1 shows the gas chromatography fingerprints of cortex Cinnamomi and rhizoma Anemarrhenae.

### 2.2 Gas chromatography fingerprint similarity

In this paper, we study the relationship between the cold-hot nature and material composition of CHMs. A GC reflects the volatile oil ingredients of a CHM. Therefore, we have conducted an exploration to reveal the material basis of cold-hot nature on

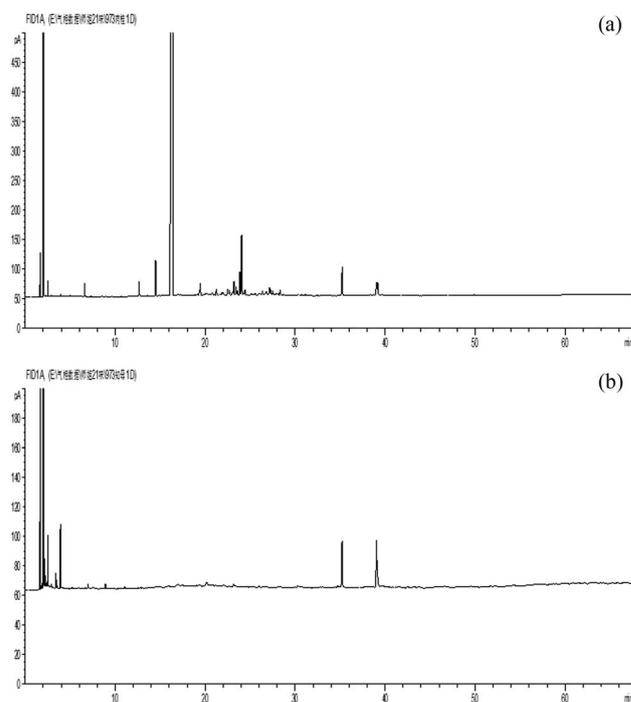


Fig. 1 GC fingerprints of cortex Cinnamomi (a) and rhizoma Anemarrhenae (b).



the basis of gas chromatography fingerprints. According to the current research, the material composition of CHMs is the material basis of their cold-hot nature.<sup>8,13</sup> If the volatile oil ingredients of CHMs are similar, we can assume that their medicinal nature is similar. Hence, CHMs with similar gas chromatography fingerprints should have the same medicinal nature.

The similarity of the chromatographic fingerprints of CHMs has been widely used in the quality evaluation of CHMs.<sup>17</sup> In this study, the similarity of chromatographic fingerprints is applied to evaluate the cold-hot nature of CHMs. We define the similarity of chromatographic fingerprints as the fingerprint similarity and semantic relevance. Fingerprint similarity is the feature similarity of the CHM ingredients, which means that the fingerprints of the two CHMs are similar. Semantic relevance depends on the cold or hot classification of CHMs, which means that if two CHMs have the same label, they are semantically similar.<sup>18</sup> We want to learn a Mahalanobis distance to measure the similarity of chromatographic fingerprints, which preserves fingerprint similarity and semantic relevance. The smaller the Mahalanobis distance is, the higher the similarity of the chromatographic fingerprints is.

**2.2.1 Distance metric learning.** The CHM chromatographic fingerprint dataset is denoted as  $X = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}$ , with  $x_i \in \mathbb{R}^d$  being the  $i$ th CHM chromatographic fingerprint in the original space and  $n$  being the total number of CHMs. The Mahalanobis distance between  $x_i$  and  $x_j$  is denoted as  $d_M(x_i, x_j)$ , which is defined as:<sup>19</sup>

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)} \quad (1)$$

In eqn (1),  $T$  denotes the transpose of a matrix or a vector, and  $M$  is a positive semi-definite matrix. If  $M = I$ ,  $d_M(x_i, x_j)$  represents a Euclidean distance. If  $M$  is restricted to be a diagonal matrix,  $d_M(x_i, x_j)$  represents a distance metric in which the different axes are given different weights. In general,  $M$  corresponds to a set of Mahalanobis distances. Because  $M$  is a positive semi-definite matrix, it can be decomposed into  $M = AA^T$ , where  $A$  is a transformation matrix. Therefore, eqn (1) can be rewritten as:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T AA^T (x_i - x_j)} = \|A^T (x_i - x_j)\| \quad (2)$$

From eqn (2), solving this Mahalanobis distance is actually equivalent to learning a transformation of Euclidean distance between CHM fingerprints in the input space. In this study, we learn transformation matrix  $A$  from the fingerprint similarity and semantic relevance. With the transformation matrix  $A$ , the Mahalanobis distance between  $x_i$  and  $x_j$  can be learned according to eqn (2).

**2.2.2 Similarity metric.** As mentioned above, fingerprint similarity is defined as the similarity of CHM chromatographic fingerprints. Fingerprint similarity reflects the similarity of the CHM ingredients. Inspired by the feature similarity of pulmonary nodule images,<sup>20,21</sup> we explored a patch alignment framework for fingerprint similarity modeling.<sup>22</sup>

Considering a given  $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ , for each  $x_i \in X$ , its local patch is  $X_i = [x_i, x_{i1}, x_{i2}, \dots, x_{ik}]^T \in \mathbb{R}^{(k+1) \times d}$ , which is

calculated by the  $k$ -nearest neighbors of  $x_i$  according to the Euclidean distance. For each patch  $X_i$ , we have a transformation model  $g_i: X_i \rightarrow Y_i$ ,  $Y_i = [y_{i1}, y_{i2}, \dots, y_{ik}]^T \in \mathbb{R}^{(k+1) \times l}$ . To learn a transformation model  $g_i$ , we would like to minimize the error between the new feature representation  $Y_i$  and the linear mapping of a patch  $X_i$ , and then align all the patches.<sup>23</sup> The local patch errors can be minimized as:

$$\min_{W_i, b_i} \| (X_i^T W_i + 1_{k+1} b_i^T) - Y_i \|^2 + \mu \| W_i \|_F^2 \quad (3)$$

In eqn (3),  $1_{k+1} \in \mathbb{R}^{k+1}$  denotes the vector of all ones;  $W_i \in \mathbb{R}^{d \times l}$  is the local projection matrix, and  $b_i \in \mathbb{R}^l$  is the bias;  $\mu$  is a regularization parameter.

It is assumed that the samples are centered, i.e.,  $X_i^T 1_{k+1} = 0$ . To obtain the optimal solution to eqn (3), we set the derivatives of the objective function with respect to  $W_i$  and  $b_i$  to zero. The solution is:

$$\begin{cases} b_i = \frac{1}{k+1} Y_i^T 1_{k+1} \\ W_i = (X_i H_{k+1} X_i^T + \mu I_d)^{-1} X_i H_{k+1} Y_i \end{cases} \quad (4)$$

where  $H_{k+1} = I_{k+1} - \frac{1}{k+1} 1_{k+1} 1_{k+1}^T$  is the local centering matrix. By substituting  $W_i$  and  $b_i$  into eqn (3) using eqn (4), eqn (3) is then rewritten as:

$$\min_{Y_i} \text{tr}(Y_i^T L_i Y_i) \quad (5)$$

In eqn (5),  $L_i = H_{k+1} - X_i^T (X_i X_i^T + \mu I_d)^{-1} X_i$ , and then, the global alignment becomes:

$$\min_Y \text{tr}(Y^T L Y) \quad (6)$$

where the global alignment matrix  $L$  is:

$$L = [S_1, \dots, S_n] \begin{bmatrix} L_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & L_n \end{bmatrix} [S_1, \dots, S_n]^T \quad (7)$$

In eqn (7),  $\{S_i\}_{i=1}^n$  is a selection matrix such that  $Y_i = S_i^T Y$ .

Define  $Y \in \mathbb{R}^{n \times l}$  is a representation of dataset  $X$  in feature space. Based on the assumption of linearization that  $Y = X^T A$ , we can obtain the errors of the global patches:

$$\min \text{tr}(A^T X L X^T A) \quad (8)$$

According to the definition of semantic relevance, it represents the separability of the cold-hot nature, which requires an increase in the class separability when the size of the within-class scatter matrix decreases or the size of the between-class scatter matrix increases. This can be modelled by the differential scatter discriminant criterion (DSDC) formula,<sup>24</sup> which is defined as:

$$A = \text{argmax}(\text{tr}(A^T S_B A) - \lambda \text{tr}(A^T S_W A)) \quad (9)$$

The variation is defined as:

$$\begin{aligned} A &= \text{argmax}(\text{tr}(A^T S_B A) - \lambda \text{tr}(A^T S_W A)) \\ &= \text{argmin}(\text{tr}(A^T (S_W - \lambda S_B) A)) \end{aligned} \quad (10)$$



In eqn (10),  $S_W$  is the within-class scatter matrix, and  $S_B$  is the between-class scatter matrix.  $\lambda$  is a nonnegative tuning parameter, which balances the relative merits of minimizing the within-class scatter with the maximization of the between-class scatter. The learned matrix  $A$  is the transformation matrix.

The similarity of CHM ingredients includes the fingerprint similarity and the semantic relevance. Therefore, we integrate eqn (8) from fingerprint similarity and eqn (10) from semantic relevance to construct the similarity metric model. The similarity metric model is as follows:

$$\begin{aligned} A &= \operatorname{argmin} \operatorname{tr}(A^T(XLX^T + S_W - \lambda S_B)A) \\ &= \operatorname{argmin} \operatorname{tr}(A^TQA) \end{aligned} \quad (11)$$

where  $Q = XLX^T + S_W - \lambda S_B$ ,  $XLX^T$  can be computed from dataset  $X$  without labels and  $S_W - \lambda S_B$  can be obtained from the dataset  $X$  with labels.

**2.2.3 Projection learning.** To learn the distance metric, it is necessary to avoid redundancy of the low-dimensional representation of the data set as much as possible. One way is to make the projection directions orthogonal, *i.e.*:

$$\begin{aligned} A^* &= \operatorname{argmin} \operatorname{tr}(A^TQA) \\ \text{s.t. } &A^T A = I \end{aligned} \quad (12)$$

In this case, the solution of optimal projections  $A^*$  can be obtained through eigenvalue decomposition on matrix  $Q$ , and  $A^*$  can be constructed by the  $u$  eigenvectors of  $M$  associated with the  $u$  smallest eigenvalues.

### 2.3 A kNN ( $k$ -nearest neighbor) scheme for cold-hot nature identification

Using the learned Mahalanobis distance, an improved similarity-based kNN scheme is developed to identify cold-hot nature. The improved kNN scheme description is shown in Fig. 2. For a CHM with an unknown nature, we first determine the ingredients of the CHM by gas chromatography, and then calculate the similarity between the gas chromatography fingerprint of the CHM and those of CHM datasets with a clear nature. The similarity-based kNN algorithm using the Mahalanobis distance (SNNM) is used to search for  $r$  similar CHM fingerprints with the smallest distances. The  $r$  'most similar' CHMs are ranked based on increasing Mahalanobis distance metrics to the query CHM. Finally, a cold nature probability is calculated to represent the degree of coldness of the CHM,

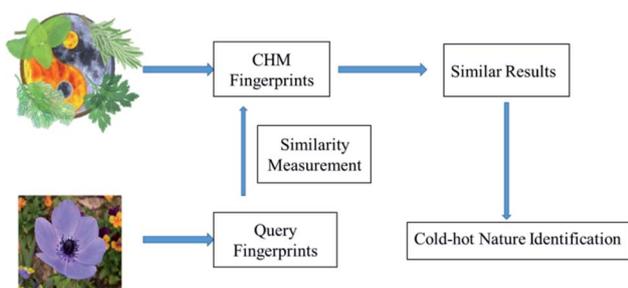


Fig. 2 A kNN scheme for identification of the nature of CHMs.

which is the ratio of the number of cold CHMs to the total number of CHMs searched. The formula is defined as eqn (13):

$$p = \frac{c}{r}, \quad c + h = r \quad (13)$$

where  $c$  is the number of cold CHMs and  $h$  is the number of hot CHMs. Given a threshold of  $P_T = 0.5$ , if  $p \geq P_T$ , we classify the queried CHM as cold; otherwise, it is hot.

### 2.4 SNNM scheme for cold-hot nature identification

The cold-hot nature identification algorithm:

Given a CHM fingerprint dataset  $X = [x_1, x_2, \dots, x_n] \in R^{d \times n}$ , and a number of classes  $C$ .

**(1) Computation of transformation matrix  $A$ .** Apply Eigenvalue decomposition on matrix  $Q$  to obtain the smallest  $r$  eigenvectors of  $Q$ . Then, according to eqn (12), calculate the transformation matrix  $A$ .

**(2) Calculation of the Mahalanobis distance  $d(x_i, x_j)$ .** According to eqn (2), calculate  $d(x_i, x_j)$  between CHM fingerprints  $x_i$  and  $x_j$ .

**(3) Similarity metric.** Replacing the Euclidean distance with the Mahalanobis distance, use the kNN ( $k$ -nearest neighbor) algorithm to calculate the  $k$  CHM fingerprints with the smallest  $k$  Mahalanobis distances.

**(4) Nature identification.** Calculate the ratio of the number of cold CHMs to the total number of CHMs.

### 2.5 Performance assessment

In this subsection, to evaluate the effectiveness and feasibility of the proposed SNNM scheme, numerous experiments are constructed for cold-hot nature identification. We compare the identification performance of our scheme with that of the state-of-the-art schemes in terms of extrapolation evaluation and stability evaluation, including information-theoretic metric learning (ITML),<sup>25</sup> large margin nearest neighbor (LMNN),<sup>20</sup> retrieval system (RS),<sup>26</sup> Pearson correlation coefficient (PCC) and ELM.<sup>27</sup> ITML and LMNN are classical distance metric learning models. RS and PCC have been used for nature identification of CHMs. ELM has been applied for the nature identification of chemical compounds in CHMs. All performance evaluation experiments are based on an existing CHM fingerprint dataset. The application allows a researcher to test a CHM of unknown nature by searching for similar CHM fingerprints with a clear nature. In this study, we first determined the CHM ingredients using gas chromatography. Secondly, we proposed a SNNM scheme for the nature identification of CHMs. Finally, we designed extensive experiments to demonstrate the feasibility of our proposed scheme.

In our experiments, the extrapolation evaluation represents the extent to which cold CHMs can be calculated on the basis of the CHMs that are retrieved based on similarity in the search. We divided the CHM fingerprint dataset into training fingerprints and test fingerprints and computed the probability of each test CHM belonging to the cold CHM group. By varying the threshold of the cold probability, a receiver operating characteristic (ROC) curve was calculated. The area under the ROC



curve (AUC) and identification accuracy (ACC) were applied to assess the performance of our scheme. The ACC formula is as follows:  $ACC = n/r$ , where  $n$  is the number of correctly identified CHMs, and  $r$  is the total number of identified CHMs. The second evaluation method, stability evaluation, reflects the proportion of retrieved CHMs that are medically relevant to the query CHMs, which can be calculated by the leave-one-CHM-out method for the whole CHM fingerprint dataset.<sup>26</sup> In this evaluation scheme, the cold probability of each CHM can be calculated from the  $r$  'most similar' CHMs in the remaining 60 test samples. Finally, the probabilities for the 61 CHMs are obtained. We calculate the stability evaluation as eqn (14):

$$R(q_i^r) = \frac{\sum_{j=1}^r I[y_i = y_j]}{r} \quad (14)$$

In eqn (14),  $R(q_i^r)$  is a function of  $r$  (the number of 'most similar' CHMs retrieved), which represents the proportion of CHMs with the same label for the  $i$ th query CHM in the first  $r$ -ranked CHMs. The overall stability evaluation is the average for all the CHMs in the test dataset. In this study, we performed extrapolation evaluation and stability evaluation with ten random experiments.

## 3. Results

### 3.1 Parameter configurations

In this section, several parameters were set for nature identification. The parameter analyses were performed based on the experiments conducted on the CHM fingerprint database. In our experiments, we analyzed several factors, namely,  $\mu$  in eqn (3) for patch building; the tradeoff parameter  $\lambda$  in eqn (9); and the number of retrieved CHMs  $k$  in the kNN scheme.

In this work, the stability evaluation was applied to set the parameters of our identification model. AUC and ACC values were calculated to evaluate the performance of our identification scheme with different parameters ( $\mu$ ,  $\lambda$ , and  $k$ ). The AUC and ACC values were computed as a function of the set

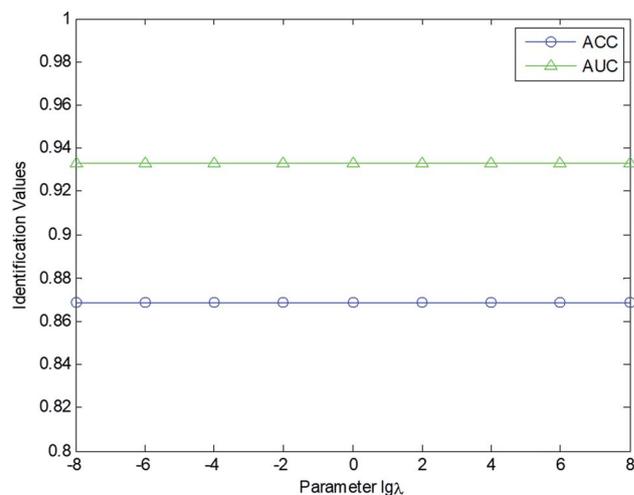


Fig. 4 Curves of AUC and ACC values with different  $\lambda$ .

parameters to obtain more comprehensive curves for evaluating the performance of our scheme. Fig. 3 shows the AUC and ACC value curves for the nature identification of the gas chromatography fingerprints using different  $\mu$  values in the range  $[10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6, 10^8]$ . As shown in Fig. 3, our scheme is not suitable for a large  $\mu$ . When  $\mu \geq 10^4$ , the identification performance of our model drops. Based on comprehensive analysis of the ACC and AUC curves, our scheme reaches the optimal value when  $\mu = 10^{-4}$ . In this experiment, the tradeoff parameter  $\lambda$  was set as 1, and the number of retrieved CHMs  $k$  was set as 5.

In this study, the effect of the tradeoff parameter  $\lambda$  in eqn (9) was investigated to evaluate the identification performance of the cold-hot nature. The value of the parameter  $\lambda$  was set within the range  $[10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6, 10^8]$ . Fig. 4 displays the AUC and ACC value curves with different  $\lambda$  values. According to Fig. 4, our scheme and fingerprint data are not sensitive to the parameter  $\lambda$ . No matter what the parameter value is, the performance of the model remains at a certain

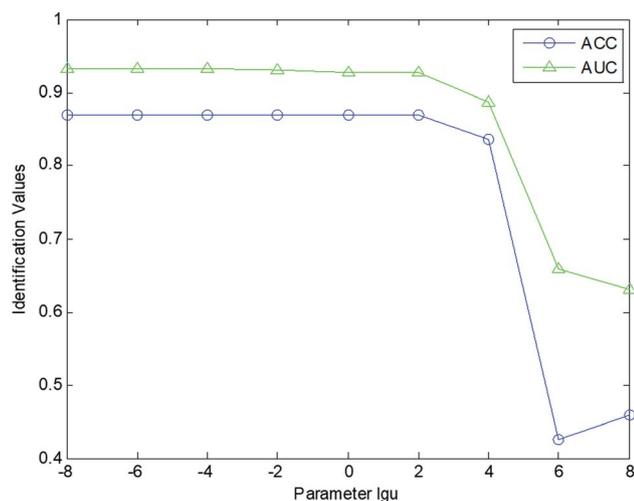


Fig. 3 Curves of the AUC and ACC values for the nature classification.

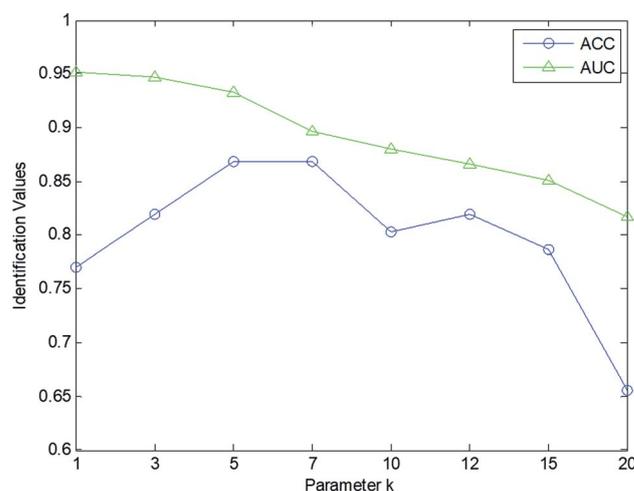


Fig. 5 Curves of AUC and ACC values with different  $k$ .



Table 1 Comparison of extrapolation evaluation

Classifier	AUC	ACC
ITML	0.872	0.823
LMNN	0.855	0.786
ELM	0.587	0.525
RS	0.882	0.824
PCC	0.834	0.754
<b>SNNM</b>	<b>0.891</b>	<b>0.852</b>

Table 2 Comparison of stability evaluation

Classifier	AUC	ACC
ITML	0.896	0.869
LMNN	0.894	0.869
ELM	0.683	0.623
RS	0.872	0.820
PCC	0.603	0.557
<b>SNNM</b>	<b>0.9333</b>	<b>0.869</b>

level. Setting  $\lambda = 1$ , the AUC and ACC values of our model are 0.8689 and 0.9333, respectively. In this experiment, the parameter  $\mu$  was set as  $10^{-4}$ , and the number of retrieved CHMs  $k$  was set as 5.

Furthermore, the parameter  $k$  in the kNN scheme was varied to evaluate the predictive performance of our scheme. The value of  $k$  was set within the range [1, 3, 5, 7, 12, 15, 20]. Fig. 5 shows the AUC and ACC value curves with different  $k$  values. From this figure, we can see that the performance of our scheme tends to decline with increasing  $k$  value. Based on comprehensive analysis of the ACC and AUC curves, our scheme reaches the optimal value when  $k = 5$ . In this experiment, the parameter  $\mu$  was set as  $10^{-4}$ , and the tradeoff parameter  $\lambda$  was set as 1.

### 3.2 Model performance assessment

To demonstrate the feasibility and the efficiency of our proposed SNNM scheme for predicting the cold-hot nature of CHMs, this study compares the prediction performance of our model with those of some classical distance metric learning algorithms (*i.e.*, ITML and LMNN) or classifiers used in CHM

nature identification (RS, PCC, and ELM). The Pearson correlation coefficient (PCC) is applied as a comparative reference to measure the similarity of gas chromatography. Table 1 displays a comparison between the performance of the extrapolation evaluation of SNNM and other algorithms. For the extrapolation evaluation, 40 CHMs were randomly selected as the training dataset, among which the number of cold and hot CHMs was about 20 each; the remaining CHMs served as the test dataset. The parameters of our model for the extrapolation evaluation were  $\lambda = 1$ ,  $\mu = 10^{-4}$ ,  $k = 7$ . The parameters of the compared models were optimized. Based on the results for the classification of cold-hot nature, we drew the following conclusions. First, the identification performance of our SNNM scheme was the best, and was better than that of the RS algorithm. The RS algorithm only considers the semantic relevance between fingerprints. Therefore, fingerprint similarity is also important for similarity metric of CHM ingredients. Secondly, distance metric learning algorithms have better classification accuracy than ELM and PCC. This demonstrates that it is feasibility for nature prediction with similarity metric of CHM ingredients, and also illustrates that CHMs with similar gas chromatography have similar medicinal nature. Thirdly, ELM with gas chromatography is poor in identifying cold-hot nature. Finally, according to the extrapolation evaluation, our scheme is efficiency.

Table 2 shows a performance comparison between the stability evaluation of SNNM and the other algorithms. Based on the cold-hot nature prediction results, we can draw similar conclusions to those obtained from Table 1. First, our SNNM scheme performs best in the identification of cold-hot nature. Secondly, distance metric learning algorithms outperform ELM and PCC in identifying the cold-hot nature. Thirdly, the stability assessment of our scheme is the best. Finally, Tables 1 and 2 comprehensively verify the feasibility and effectiveness of our scheme.

### 3.3 Nature identification examples

The leave-one-CHM-out method was applied to calculate examples of the nature identification. The nature identification examples of the two CHMs are displayed in Table 3. The query CHMs (the second row) and their top  $k = 7$  similar reference CHMs are shown in this table. The similar reference CHMs were

Table 3 Nature identification examples. The top  $k = 7$  similar CHMs are arranged in the order of monotonically increasing Mahalanobis distance. Cold/hot nature labels are denoted in brackets

Prediction example	CHMs with a cold nature	CHMs with a hot nature
Query CHM	<i>Anemarrhena asphodeloides</i> Bunge (cold)	<i>Euodiae fructus</i> (hot)
Similar reference CHMs	<i>Phellodendri chinensis cortex</i> (cold)	<i>Notopterygii rhizoma et radix</i> (hot)
	<i>Isatidis folium</i> (cold)	<i>Corydalis rhizoma</i> (hot)
	<i>Lophatheri herba</i> (cold)	<i>Aconiti lateralis radix praeparata</i> (hot)
	<i>Stephaniae tetrandrae radix</i> (cold)	<i>Alpiniae katsumadai semen</i> (hot)
	<i>Puerariae lobatae radix</i> (cold)	<i>Psoraleae fructus</i> (hot)
	<i>Gardeniae fructus</i> (cold)	<i>Nardostachyos radix et rhizoma</i> (hot)
	<i>Notopterygii rhizoma et radix</i> (hot)	<i>Aucklandiae radix</i> (hot)



Table 4 Confusion matrix of the 61 CHMs

Ground truth	Identification	
	Cold	Hot
Cold	26	4
Hot	4	27

calculated using SNNM and arranged in order of monotonically increasing Mahalanobis distance.

A cold CHM (*Anemarrhena asphodeloides* Bunge) and a hot CHM (*Euodiae fructus*) were selected as examples to explain the principle of cold-hot nature identification. In the second column, the query CHM is *Anemarrhena asphodeloides* Bunge, which is a CHM with a cold nature. The reference CHMs similar to it are six CHMs with a cold nature, and one CHM with a hot nature. The calculated cold nature probability is 85.7%, which means that the query CHMs are more likely to have a cold nature. In the third column, the query CHM is *Euodiae fructus*. The calculated reference CHMs are all hot-nature CHMs. Its cold nature probability is 0, meaning that the query CHM is most likely to have a hot nature. The identification examples indicate that similar gas fingerprints can represent the same CHM nature.

### 3.4 Overall prediction performance

In this paper, we performed a holistic evaluation of the proposed SNNM method. A confusion matrix, recall, precision and *F*-score were introduced as evaluation indexes. The identification confusion matrix of 61 CHMs is displayed in Table 4. The identification accuracy of hot-nature CHMs is 87.1% (27/31), while the prediction accuracy of cold-nature CHMs is 86.7% (26/30). The total identification accuracy is 86.9% (53/61). The recall, precision and *F*-score of nature identification of the 61 CHMs are listed in Table 5. Based on comprehensive analysis of Tables 4 and 5, we can conclude that our method has good prediction performance for cold-hot nature using gas chromatography. The ingredients of the volatile oils of CHMs are closely related to their cold-hot nature. Fig. 6 displays the ROC curve of the cold-hot nature identification.

## 4. Discussion

CHM gas chromatography fingerprints pose a challenge to existing distance metric methods. Traditional similarity measurement methods, such as Pearson correlation, suffer from high-dimensional problems. Distance metric learning

Table 5 Recall, precision and *F*-score of the 61 CHMs

	Cold	Hot
Recall	86.7%	87.1%
Precision	86.7%	87.1%
<i>F</i> -Score	86.7%	87.1%

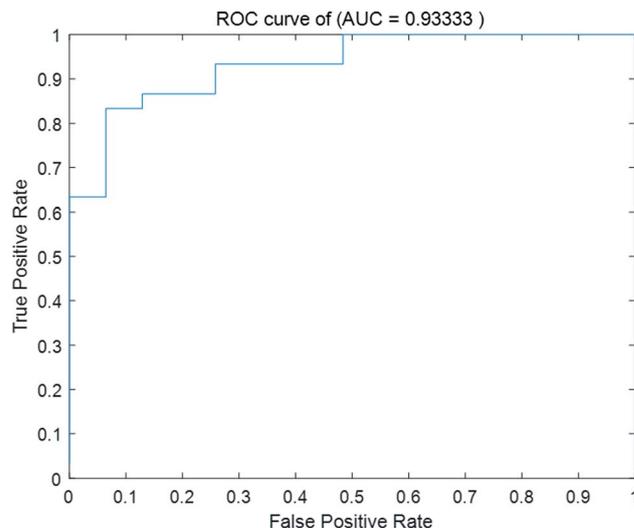


Fig. 6 ROC curve of cold-hot nature identification.

methods, such as ITML and LMNN, only consider the semantic relevance of the components without including the fingerprint similarity. However, semantic relevance alone does not reflect all the similarities of CHMs. Herein, we introduce the fingerprint similarity and semantic relevance to represent the similarity of the CHM ingredients. The model is more consistent with the characteristics of the fingerprint data. We find that fingerprint similarity is very important in similarity measurement, which can effectively improve the identification accuracy.

Volatile ingredients are an important part of CHMs and play an important role in the therapeutic effect of traditional Chinese medicine. In this study, the volatile ingredients of CHMs are extracted *via* gas chromatography. It is found that the volatile ingredients are closely related to the cold-hot nature of the CHMs. According to our previous studies on the nature identification of CHMs, gas chromatography has a better nature identification rate than UV spectroscopy. This indirectly proved the correlation between the volatile oils and the cold-hot nature of CHMs.

Classical classification approaches, such as ELM, are general classifiers that do not consider the characteristics of the data. These methods suffer from the problem of small samples and high dimensionality, resulting in low classification accuracy. Compared to classical classification approaches, our scheme not only considers the class separability of the samples, but also introduces the fingerprint similarity. Therefore, our scheme achieves better identification performance.

However, our research still has some limitations. First, this study only analyzes volatile oil ingredients *via* gas chromatography. Other ingredients in the CHMs are not taken into account in this study. In the future, we want to explore cold-hot nature identification of CHMs based on total component information. Second, we studied the similarity of the gas chromatography data using a distance metric. The fingerprints have the characteristics of high dimension and small sample. Based on these characteristics, the design of the forecasting



model will be the focus in the future. Thirdly, this study focuses on modeling a similar scheme for cold-hot nature identification. The GC characteristics have not been thoroughly mined. In the future, we will combine more effective fingerprint information to represent CHM ingredients for cold-hot nature identification.

## 5. Conclusions

In this study, a SNNM scheme is proposed to predict the cold-hot nature of CHMs. Gas chromatography is used to analyze the volatile oil ingredients of CHMs. It is found that the volatile oil ingredients of CHMs are closely related to their cold-hot nature. We demonstrate that if the ingredients of CHMs are similar, their nature is similar. Based on the gas chromatography fingerprints of the CHMs, effective experiments demonstrate that our scheme has better performance than classical classifiers in identifying the cold-hot nature of the CHMs.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work was supported by the National Key Basic Research Development Program (973 Program) (No. 2007CB512600), National Natural Science Foundation of China (No. 81473369), Key Research and Development Plan of Shandong Province (No. 2016CYJS08A01-1), and Shandong Province TCM Science and Technology Development Plan Project (2019-0037).

## Notes and references

- 1 J. Gao and C. Chen, *Acta Univ. Tradit. Med. Sin. Pharmacol. Shanghai*, 2007, **21**, 16–18.
- 2 B. Ouyang, Z. G. Wang and P. Wang, *J. Beijing Univ. Tradit. Chin. Med.*, 2006, **29**, 592–594.
- 3 L. J. Wu, Y. Y. Wang, Y. D. Li, X. G. Zhang, S. Li and Z. Q. Zhang, *IEE Proc.: Syst. Biol.*, 2007, **1**, 51–60.
- 4 R. Jin, B. Zhang, X.-Q. Liu, S.-M. Liu, X. Liu, L.-Z. Li, Q. Zhang and C.-M. Xue, *J. Chin. Integr. Med.*, 2011, **9**, 715–724.
- 5 H.-Y. Wan, X.-Y. Kong, X.-M. Li, H.-W. Zhu, X.-H. Su and N. Lin, *China J. Chin. Mater. Med.*, 2014, **39**, 3813–3818.
- 6 Y. Zhao, L. Jia, J. Wang, W. Zou, H. Yang and X. Xiao, *Pharm. Biol.*, 2016, **54**, 1298–1302.
- 7 F. Liang, L. Li, M. Wang, X. Niu, J. Zhan, X. He, C. Yu, M. Jiang and A. Lu, *J. Ethnopharmacol.*, 2013, **148**, 770–779.
- 8 X. Fu, L. H. Mervin, X. Li, H. Yu and J. Li, *J. Chem. Inf. Model.*, 2017, **57**, 468–483.
- 9 G. J. Xu, J. H. Hu and W. Yang, *J. Nanjing Pharmaceut. College*, 1961, **6**, 92–100.
- 10 Q. D. Jiang, W. G. Yang, H. Cai, M. Ma, H. Zhang, P. Liu, J. Chen and J. Duan, *China J. Chin. Mater. Med.*, 2016, **41**, 2500–2505.
- 11 W.-H. Liu, Y. Li, Y.-J. Ji, P. Wang, Y.-Q. Zhang and F.-Z. Xue, *J. Shandong Univ., Health Sci.*, 2012, **50**, 151–154.
- 12 X.-X. Zhang, Y. Li, Y.-J. Ji, P. Wang, Y.-Q. Zhang and F.-Z. Xue, *J. Shandong Univ., Health Sci.*, 2012, **50**, 143–146.
- 13 G. H. Wei, X. J. Fu and Z. G. Wang, *J. Chem. Inf. Model.*, 2019, **59**, 5065–5073.
- 14 B. Nie, Z.-L. Hao, B. Gui, Z. Wang, J.-Q. Du, G.-L. Wang and X. Zhang, *Journal of Jiangxi University of Traditional Chinese Medicine*, 2015, **27**, 82–86.
- 15 W. Long, P. Liu, J. Xiang, X. Pi, J. Zhang and Z. Zou, *Comput. Methods Programs Biomed.*, 2011, **101**, 253–264.
- 16 H. Li, Q. Xu, J. Zhang and H. Xu, *2017 International Conference on Medical Science and Human Health (MSHH 2017)*, 2017, pp. 174–179.
- 17 M. J. Li, Y. R. Sha, X. M. Luo, P. Y. Gong and J. Gu, *Chin. Tradit. Herb. Drugs*, 2021, **52**, 1–6.
- 18 Y. Liu, J. Rong, M. Lily, S. Rahul, G. Adam, B. Zheng, S. C. H. Hoi and S. Mahadev, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, 30–44.
- 19 K. Q. Weinberger, J. Blitzer and L. K. Saul, *J. Mach. Learn. Res.*, 2009, **10**, 207–244.
- 20 G. Wei, H. Ma, W. Qian and M. Qiu, *Curr. Med. Imaging Rev.*, 2017, **13**, 210–216.
- 21 G. Wei, H. Ma, W. Qian and M. Qiu, *Med. Phys.*, 2016, **43**, 6259–6269.
- 22 T. Zhang, D. Tao, X. Li and J. Yang, *IEEE Trans. Knowl. Data Eng.*, 2009, **21**, 1299–1313.
- 23 D. Tao, X. Li, X. Wu and S. J. Maybank, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, 1700–1715.
- 24 J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, *The International Conference on Machine Learning (Corvallis, Oregon, USA)*, 2007, pp. 209–216.
- 25 G. H. Wei, X. J. Fu and Z. G. Wang, *TMR Mod. Herb. Med.*, 2019, **2**, 183–191.
- 26 E. Malar, A. Kandaswamy, D. Chakravarthy and A. G. Dharan, *Comput. Biol. Med.*, 2012, **42**, 898–905.

