OXFORD

Briefings in Bioinformatics, 2023, 1–13
https://doi.org/10.1093/bib/bbad259
Problem Solving Protocol

Predicting metabolite–disease associations based on auto-encoder and non-negative matrix factorization

Hongyan Gao, Jianqiang Sun, Yukun Wang, Yuer Lu, Liyu Liu, Qi Zhao and Jianwei Shuai

Corresponding authors. Q. Zhao, School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. Tel:/Fax: 0086412889818; E-mail: zhaoqi@lnu.edu.cn; J.W. Shuai, Wenzhou Institute and Wenzhou Key Laboratory of Physics, University of Chinese Academy of Sciences, Wenzhou 325001, China. Tel:/Fax: 008657788017501; E-mail: jianweishuai@xmu.edu.cn

Abstract

Metabolism refers to a series of orderly chemical reactions used to maintain life activities in organisms. In healthy individuals, metabolism remains within a normal range. However, specific diseases can lead to abnormalities in the levels of certain metabolites, causing them to either increase or decrease. Detecting these deviations in metabolite levels can aid in diagnosing a disease. Traditional biological experiments often rely on a lot of manpower to do repeated experiments, which is time consuming and labor intensive. To address this issue, we develop a deep learning model based on the auto-encoder and non-negative matrix factorization named as MDA-AENMF to predict the potential associations between metabolites and diseases. We integrate a variety of similarity networks and then acquire the characteristics of both metabolites and diseases through three specific modules. First, we get the disease characteristics from the five-layer auto-encoder module. Later, in the non-negative matrix factorization module, we extract both the metabolite and disease characteristics. Furthermore, the graph attention auto-encoder module helps us obtain metabolite characteristics. After obtaining the features from three modules, these characteristics are merged into a single, comprehensive feature vector for each metabolite–disease pair. Finally, we send the corresponding feature vector and label to the multi-layer perceptron for training. The experiment demonstrates our area under the receiver operating characteristic curve of 0.975 and area under the precision–recall curve of 0.973 in 5-fold cross-validation, which are superior to those of existing state-of-the-art predictive methods. Through case studies, most of the new associations obtained by MDA-AENMF have been verified, further highlighting the reliability of MDA-AENMF in predicting the potential relationships between metabolites and diseases.

Keywords: metabolites, diseases, auto-encoder, non-negative matrix factorization, feature splicing, multi-layer perceptron

INTRODUCTION

With the advancement of medicine, there has been an increasing focus on the connection between metabolites and diseases. The metabolism is the vital process within the organism, and disease is a life process that can cause various physical symptoms, including metabolic process disorders [1]. The occurrence of diseases is accompanied by changes in metabolites. We can determine the progression of diseases through certain metabolites. For instance, fasting blood glucose, blood glucose in the second hour after meal and glycosylated hemoglobin can be checked in diabetes. If the fasting blood glucose is more than 7.0 mmol/l, the blood glucose in the second hour after meal is more than 11.1 mmol/l and the glycosylated hemoglobin is more than 6.5%, diabetes is highly probable. In addition, intestinal microbiota and its metabolites also play a crucial role in the development of many metabolic diseases, including obesity [2], nonalcoholic fatty liver [3] and cardiovascular disease [4]. Microbes in the gastrointestinal tract are abundant and can metabolize dietary nutrients into a variety of bioactive substances. The metabolic and immune potential of intestinal microbiota determines its importance in host health and disease [5]. Analyzing the relationship between maternal metabolites derived from mass spectrometry during pregnancy and congenital heart disease in offspring, it is observed that amino acid metabolism during pregnancy, androgen steroid lipid and succinyl carnitine level may be significant contributing factors to coronary heart disease [6]. In addition, piperine and its metabolites may be potential substances for the treatment of heart and liver diseases, chronic inflammation, neurodegenerative diseases and cancer [7]. Moreover, in identifying metabolites related to the etiology of Alzheimer's disease, glutamine and free cholesterol in ultra-high density lipoprotein have a protective effect on Alzheimer's disease [8].

It is a meaningful endeavor to devote to metabonomics to study the pathogenesis of diseases, but traditional biological experiment methods can be time consuming and laborious, and may not yield optimal results. During recent years, many kinds of

Hongyan Gao is a graduate student in University of Science and Technology Liaoning. Her research interests include bioinformatics and deep learning.
Jianqiang Sun is an associate professor in Linyi University. His research interests include bioinformatics and deep learning.
Yukun Wang is an associate professor in University of Science and Technology Liaoning. His research interests include machine learning and QSAR modeling.
Yuer Lu is a research assistant in Wenzhou Institute, University of Chinese Academy of Sciences. Her research interests include deep learning and bioinformatics.
Liyu Liu is a professor in Chongqing University. His main research interests include the application of swarm robotics in the field of active matter, as well as interdisciplinary research on the fusion of biointelligence, artificial intelligence and mechanical intelligence.
Qi Zhao is a professor in University of Science and Technology Liaoning. His research interests include bioinformatics, complex network and machine learning.
Jianwei Shuai is a professor and the director of biomedical physics center in Wenzhou Institute, University of Chinese Academy of Sciences. His research interests include biophysics, deep learning and bioinformatics.

Received: April 25, 2023. **Revised:** June 18, 2023. **Accepted:** June 27, 2023

researches such as computational toxicology [9], miRNA–lncRNA interaction prediction [10–12], miRNA–disease association prediction [13–15] and circRNA–disease association prediction [16–18] have been carried out in bioinformatics. These studies have promoted the development of computational methods for predicting metabolite–disease associations to a certain extent. For example, in 2018, Hu *et al.* used random walk to identify disease-related metabolites [19]. Later, Lei *et al.* developed a calculation method based on KATZ to predict the metabolite–disease associations [20]. This is also the first application of KATZ algorithm in the field of metabolomics. In 2020, Lei *et al.* proposed a linear neighborhood similarity with improved bipartite network projection algorithm to predict associations between metabolites and diseases [21]. Furthermore, Zhao *et al.* introduced a Deep-DRM model in 2021 [22], employing graph convolution networks and principal component analysis (PCA) for identifying metabolite–disease associations. In the same year, Zhang *et al.* proposed LGBMMDA [23], a method that extracts features from various measurements, applies PCA for noise reduction and utilizes the LGBM classifier for analysis. In 2022, Tie *et al.* proposed a metabolite–disease association prediction algorithm based on DeepWalk and random forest [24]. During the same period, Sun *et al.* came up with a graph neural network with attention mechanisms to predict the associations between metabolites and diseases [25]. Although many computational models have been proposed to predict the potential associations between metabolites and diseases, the application of deep learning methods is relatively few, and the accuracy of the predictive results still requires further improvement.

Based on the situation described previously, we develop a powerful deep learning method called MDA-AENMF to accurately predict the relationships between metabolites and diseases. We integrate multiple similarity networks of diseases or metabolites and apply three distinct modules to acquire meaningful features respectively. First, the five-layer auto-encoder module is applied to obtain the features of diseases. Second, non-negative matrix factorization (NMF) module is used to extract the features of both metabolites and diseases. Third, we employ the graph attention auto-encoder (GAE) module to get the features of metabolites. The features obtained from three module are then combined into a long feature vector for each metabolite–disease pair, and this feature vector and corresponding tag are sent to the multi-layer perceptron (MLP) classifier for training. To evaluate the performance of MDA-AENMF, we employ 5-fold cross-validation (5-fold CV) and compare our results with those of five state-of-the-art models. Our results demonstrate that MDA-AENMF outperforms these other models in terms of area under the receiver operating characteristic curve (AUC). Moreover, we conduct case studies based on MDA-AENMF, and it is found that most of the predicted top 20 metabolite disease pairs (MDPs) are verified, which further demonstrates the reliability and superiority of MDA-AENMF in predicting potential metabolite–disease associations.

MATERIALS AND METHODS

Datasets

Human Metabolome Database (HMDB, https://hmdb.ca/) is the most comprehensive database on biologically specific metabolism. There are 4536 metabolite–disease pairs sourced from HMDB in our datasets, including 2262 metabolites and 216 diseases. The selected 216 diseases such as uremia, leukemia and hepatitis are common diseases in our life.

We transform these associations into an adjacency matrix $A\,(n_m * n_d)$ to describe the relationships between metabolites and diseases. n_m represents the amount of metabolites and $n_d$is the number of diseases. If metabolite m_i is associated with disease d_j, then the value of $A(i, j)$ is equal to 1, otherwise 0.

MDA-AENMF

The workflow chart of MDA-AENMF is presented in Figure 1. The process begins with the construction of similarity networks for diseases and metabolites, which are then integrated using a non-linear approach. After that, we employ three modules to perform feature extraction. We obtain disease characteristics from five-layer auto-encoder module. In NMF module, we acquire both metabolites and disease characteristics. Finally, GAE module is used to extract metabolite characteristics. These features are combined for each metabolite–disease pair into a feature vector, which is then sent, along with the corresponding label, to MLP classifier for training.

Similarity network construction

Disease semantic similarity

In order to measure the semantic similarity between diseases, we utilize a directed acyclic graph (DAG) constructed from the Medical Subject Headings (MeSH) as descriptors of diseases. Those descriptors can be obtained from MeSH database of the National Medical Library, which provides a comprehensive resource for medical terminology (https://meshb.nlm.nih.gov/). The constructed DAG of disease d can be expressed as DAG(d) = (d, T(d), E(d)), where d represents disease d, T(d) is a disease set including disease d and its ancestors, and E(d) manifests a set of edges of disease d. The semantic contribution of disease n to disease d in DAG(d) can be calculated as

$$\begin{cases} D_d(n) = 1 & \text{if } n = d \\ D_d(n) = \max\left\{\varDelta * D_d\left(n'\right) | n' \in \text{children of } d\right\} & \text{if } n \neq d \end{cases} \tag{1}$$

where Δ is the semantic contribution decay factor, which is generally set to 0.5 [26], and disease $n \in \mathrm{T}(d)$.

Then, we sum the contribution value of each ancestor with the contribution value of the disease itself to get the semantic score for disease d as follows:

$$DV(d) = \sum_{n \in \mathrm{N}(\mathbf{d})} D_d(n) \tag{2}$$

Our analysis reveals that diseases that share a larger proportion of DAGs tend to exhibit greater similarity. Therefore, we can calculate the disease semantic similarity (DSS) between disease d_i and disease d_j by the following formula:

$$\mathrm{DSS}\left(d_i, d_j\right) = \frac{\sum\limits_{t \in T(\mathrm{i}) \cap T(\mathrm{j})} D_{\mathrm{i}}(\mathrm{t}) + D_j(\mathrm{t})}{DV(\mathrm{i}) + DV(j)} \tag{3}$$

Metabolite structural similarity

Chemical properties are the most important properties for metabolites to participate in biochemical reactions. Each metabolite has its own unique chemical structure that can be described using the simplified molecular input line entry system (SMILES), an internationally recognized notation. To

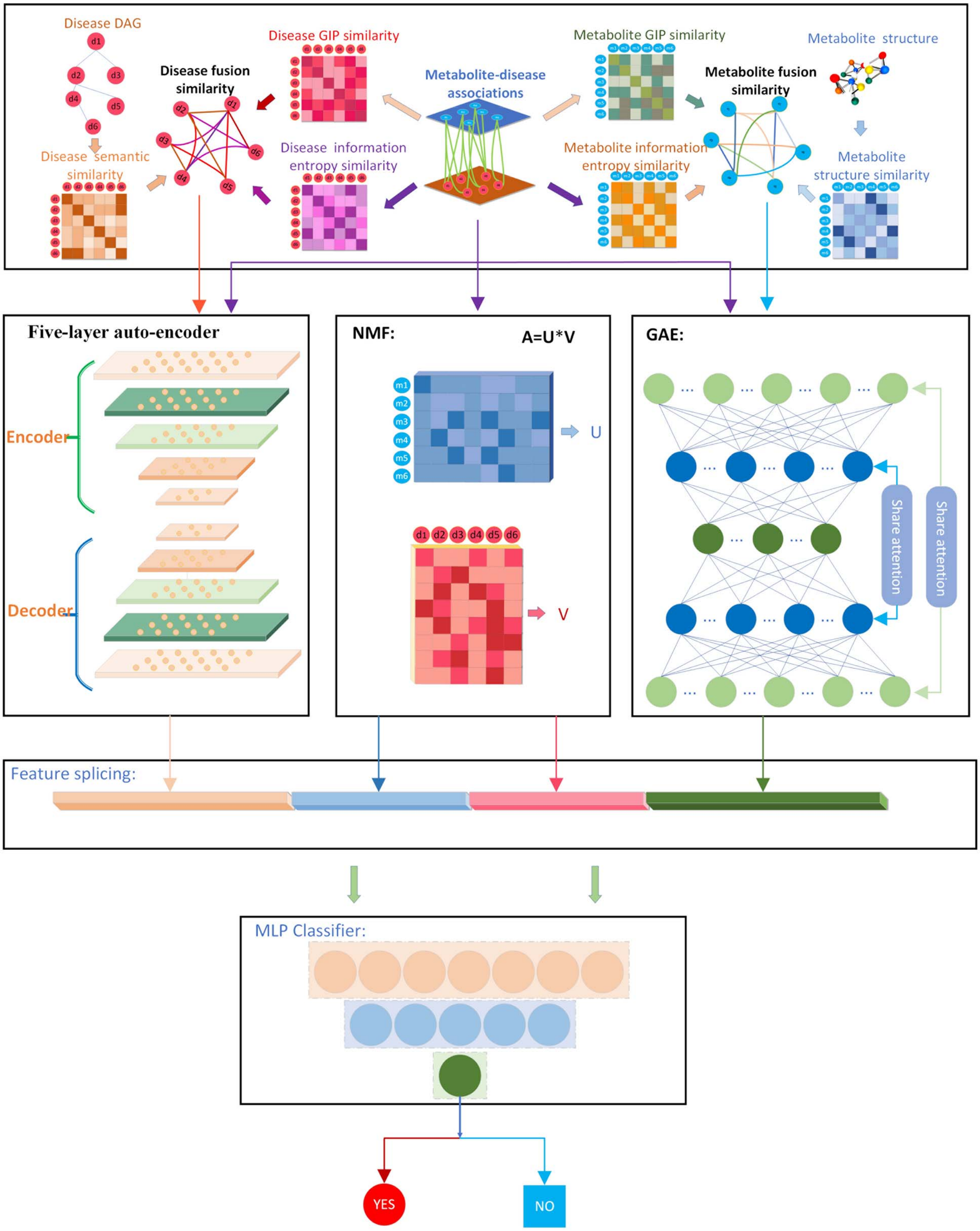
Disease DAG
d1
d2
d3
d4
d5
d6
Disease semantic similarity
Disease fusion similarity
Disease GIP similarity
Disease information entropy similarity
Metabolite-disease associations
Metabolite GIP similarity
Metabolite fusion similarity
Metabolite structure
Metabolite information entropy similarity
Metabolite structure similarity
m1
m2
m3
m4
m5
m6
Five-layer auto-encoder
Encoder
Decoder
NMF:
A=U*V
U
V
GAE:
Share attention
Share attention
Feature splicing:
MLP Classifier:
YES
NO

Figure 1. The workflow of MDA-AENMF.

obtain information about the structure of metabolites, we input SMILES of the metabolites into the software PaDEL-Descriptor to calculate the descriptors and fingerprints of the metabolites. Molecular descriptors are symbols used to describe the structural or experimental information about chemical molecules. 1D and 2D descriptors provide information about chemical composition and topology, respectively. Pubchem fingerprint is a widely used chemical molecular fingerprint for drug screening and similarity searching, which has high information richness and sensitivity. This fingerprint utilizes 881 fixed-length substructures that can cover most functional groups and ring structures in chemical molecules. Each substructure is encoded as a binary digit, where

1 represents its existence and 0 represents its absence. By using PaDEL-Descriptor, each metabolite can be assigned a 2209 dimensional vector, consisting of 1328 for 1D and 2D descriptors and 881 for fingerprints, to describe its chemical properties. Since the dimensions have different scales, we standardize each dimension. The standardization of Z score [22] is introduced as follows:

$$\hat{m}_i^k = \frac{m_i^k - \mathbf{mean}\left(m^k\right)}{\mathbf{std}\left(m^k\right)} \tag{4}$$

Here, m_i denotes the metabolite i and m^k is the vector made up of kth dimension of all metabolites. m_i^k represents the kth dimension of metabolite i. $\hat{m}_i^k$ is the value of m_i^k after standardization.

Through the characteristic vectors of these metabolites, we can calculate metabolite structural similarity (MSS) between metabolite m_i and metabolite m_j.

$$\mathrm{MSS}\left(\hat{m}_i, \hat{m}_j\right) = \frac{\sum_{k=1}^{2209} \hat{m}_i^k \times \hat{m}_j^k}{\sqrt{\sum_{k=1}^{2209}\left(\hat{m}_i^k\right)^2} \times \sqrt{\sum_{k=1}^{2209}\left(\hat{m}_j^k\right)^2}} \tag{5}$$

Disease or metabolite Gaussian kernel similarity

The Gaussian kernel similarity of diseases is based on the Gaussian kernel function to calculate the similarity score between two diseases. In the network of diseases, diseases that are similar in nature tend to be closely associated with each other. We use the following Gaussian kernel formula to compute the disease Gaussian interaction profile kernel similarity (DGIP) for disease d_i and disease d_j.

$$\mathrm{DGIP}\left(d_i, d_j\right) = \mathbf{exp}\left(-\omega_d \left\| IP\left(d_i\right) - IP\left(d_j\right)\right\|^2\right) \tag{6}$$

$$\omega_d = \frac{{\omega_d}' * n_d}{\sum_{i=1}^{n_d} \mathrm{IP}\left(d_i\right)^2} \tag{7}$$

Here, n_d represents the number of diseases, and $\mathrm{IP}(d_i)$ and $\mathrm{IP}(d_j)$ refer to the vectors related to the disease d_i and disease d_j, respectively. ω_d is the normalized kernel bandwidth, which determines the density of samples in the feature space and controls the width of the similarity function in terms of similarity. Selecting an appropriate bandwidth value can achieve a balance between the smoothness and sharpness of the similarity function, which leads to better classification results. ω_d can be updated by the new normalized bandwidth ω_d'. ω_d' usually is set to 1.

Similarly, we also calculate the metabolite Gaussian interaction profile kernel similarity (MGIP) for metabolite m_i and metabolite m_j by the above approach as follows:

$$\mathrm{MGIP}\left(m_i, m_j\right) = \mathbf{exp}\left(-\omega_m \left\| IP\left(m_i\right) - IP\left(m_j\right)\right\|^2\right) \tag{8}$$

$$\omega_m = \frac{{\omega_m}' * n_m}{\sum_{i=1}^{n_m} \mathrm{IP}(m_i)^2} \tag{9}$$

where n_m indicates the quantity of metabolites, $\mathrm{IP}(m_i)$ and $\mathrm{IP}(m_j)$ denote the vectors related to the metabolite m_i and metabolite m_j, respectively. ω_m is the normalized kernel bandwidth, which can be updated by the new normalized bandwidth ω_m'. ω_m' is usually set to 1.

Disease or metabolite similarity based on information entropy

In 1948, Shannon introduced the concept of entropy in thermodynamics into information theory and proposed the concept of 'information entropy'. In the previous study, information entropy was also used to calculate the similarity between miRNA and disease [27]. In this work, we utilize information entropy and common information from both metabolites and diseases to determine the disease or metabolite similarity. Taking the calculation of metabolite similarity based on information entropy as an example: T_m^A represents the disease set associated with metabolite A, T_m^A={$T_m^A(1)$, $T_m^A(2)$, …, $T_m^A(n_{na})$}. Similarly, T_m^B denotes the disease set associated with metabolite B, T_m^B={$T_m^B(1)$, $T_m^B(2)$, …, $T_m^B(n_{nb})$}. n_{na} and n_{nb} correspond to the number of diseases associated with metabolite A and metabolite B, respectively.

$$H\left(T_{\mathbf{m}}^{A}\right) = -\sum_{\mathrm{i}=1}^{n_{ma}} p\left(T_{\mathbf{m}}^{A}(\mathrm{i})\right) \log_2\left(p\left(T_{\mathbf{m}}^{A}(\mathrm{i})\right)\right) \tag{10}$$

$$\mathrm{p}\left(T_m^A(\mathrm{i})\right) = \frac{n\left(T_m^A(\mathrm{i})\right)}{N} \tag{11}$$

N is the correlation number of all metabolites and diseases, $n(T_m^A(\mathrm{i}))$ is the count of known associations between the ith disease and all metabolites in the disease set linked with metabolite A. The metabolite similarity based on information entropy (MSIE) between metabolite A and metabolite B is calculated as follows:

$$\mathrm{MSIE}\,(A, B) = \frac{2 \times H\left(T_m^A \cap T_m^B\right)}{H\left(T_m^A\right) + H\left(T_m^B\right)} \tag{12}$$

$T_m^A \cap T_m^B$ refers to the associated diseases shared by metabolite A and metabolite B.

Similarly, we can get the disease similarity based on information entropy (DSIE) by computing information entropy and mutual information from diseases and metabolites.

Integration of similarity networks for metabolites or diseases

In contrast to the traditional linear integration method [12], we employ a nonlinear approach to integrate distinct similarity networks for metabolites or diseases [28], which could better capture the shared and complementary information of each data source and reduce noise. Also, these noises may originate from the data itself, such as measurement errors and sample collection bias, or from the feature extraction process, such as bias in feature selection or noise introduced by feature reduction. After integration, we obtain a comprehensive similarity network for either metabolites or diseases. In this subsection, we illustrate the integration process for metabolite similarity networks in details.

To begin, we need to normalize each network of metabolites. When we take MSS as an example, the normalization process is as follows:

$$\mathbf{SM}_{\mathbf{MSS}}\left(i, j\right) = \begin{cases} \frac{\mathrm{MSS}(i,j)}{2 * \sum_{k \neq i} \mathrm{MSS}(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases} \tag{13}$$

All diagonal elements are set to 1/2, and the sum of elements in each line is equal to 1. We obtain $\mathbf{SM}_{\mathbf{MSS}}$ after performing normalization of MSS. We can also acquire $\mathbf{SM}_{\mathbf{MGIP}}$ and $\mathbf{SM}_{\mathbf{MSIE}}$ in the same way. Then, we employ K nearest neighbors (KNNs) to compute the local affinity $\mathbf{S\_kn}_{\mathbf{MSS}}$ for MSS between metabolite i

and metabolite j as follows:

$$\mathbf{S\_kn}_{\mathrm{MSS}}\left(i,j\right)=\begin{cases}\frac{\mathrm{MSS}(i,j)}{\sum_{k\in N_i}\mathrm{MSS}(i,k)}, & j\in N_i\\ 0, & \text{otherwise}\end{cases}\tag{14}$$

N_i is the set of KNNs of the given node, where N_i is equal to the total number of metabolites divided by 10. This operation is based on the principle that closer distances imply higher similarity. To ensure accuracy, we assign a similarity score of 0 to remote nodes that are far away from the given node. Similarly, we can get $\mathbf{S\_kn}_{\mathrm{MGIP}}$ and $\mathbf{S\_kn}_{\mathrm{MSIE}}$ by the above way.

Then, we iterate to update each of similarity networks as below:

$$\mathbf{SM}_{\mathrm{MSS}}^{(t)}=\mathbf{S\_kn}_{\mathrm{MSS}}\times\left(\frac{\sum_{k\neq\mathrm{MSS}}\mathbf{SM}_k^{(t-1)}}{m-1}\right)\times\left(\mathbf{S\_kn}_{\mathrm{MSS}}\right)^T\tag{15}$$

$$\mathbf{SM}_{\mathrm{MGIP}}^{(t)}=\mathbf{S\_kn}_{\mathrm{MGIP}}\times\left(\frac{\sum_{k\neq\mathrm{MGIP}}\mathbf{SM}_k^{(t-1)}}{m-1}\right)\times\left(\mathbf{S\_kn}_{\mathrm{MGIP}}\right)^T\tag{16}$$

$$\mathbf{SM}_{\mathrm{MSIE}}^{(t)}=\mathbf{S\_kn}_{\mathrm{MSIE}}\times\left(\frac{\sum_{k\neq\mathrm{MSIE}}\mathbf{SM}_k^{(t-1)}}{m-1}\right)\times\left(\mathbf{S\_kn}_{\mathrm{MSIE}}\right)^T\tag{17}$$

Here, m represents the number of distinct similarity networks of metabolites. There are three kinds of metabolite similarity networks, so m is equal to 3. k represents the current selected metabolite similarity network, which can take values of MSS or MGIP or MSIE. t refers to the number of iterations performed. $\mathbf{SM}_{\mathrm{MSS}}^{(t)}$, $\mathbf{SM}_{\mathrm{MGIP}}^{(t)}$ and $\mathbf{SM}_{\mathrm{MSIE}}^{(t)}$ denote the status matrix of $\mathbf{SM}_{\mathrm{MSS}}$, $\mathbf{SM}_{\mathrm{MGIP}}$ and $\mathbf{SM}_{\mathrm{MSIE}}$ after t iterations, respectively. Renormalizing and calculating the comprehensive similarity network is needed after each iteration. The comprehensive similarity network SM is indicated as follows:

$$\mathbf{SM}=\frac{\mathbf{SM}_{\mathrm{MSS}}^{(t)}+\mathbf{SM}_{\mathrm{MGIP}}^{(t)}+\mathbf{SM}_{\mathrm{MSIE}}^{(t)}}{3}\tag{18}$$

Once the condition $\frac{\left\|\mathbf{SM}_{\mathrm{k}}^{(t)}-\mathbf{SM}_k^{(t-1)}\right\|}{\left\|\mathbf{SM}_k^{(t-1)}\right\|}<10^{-6}$ is met, the iteration is terminated, and then we convert SM into a symmetric matrix through $\mathbf{SM}'=\frac{\mathrm{SM}+\mathrm{SM}^T}{2}$. The final symmetric matrix $\mathbf{SM}'$ represents the outcome of our integration. In the similar way, we can also apply the above rule to get the disease integration network $\mathbf{SD}'$.

Features extraction process

Disease features extraction from five-layer auto-encoder

Auto-encoder is composed of two components, an encoder and a decoder. Encoder maps data from D dimensions to M dimensions, while decoder maps data from M dimensions back to D dimensions. Encoding can be viewed as a process of compressing data, while decoding is the process of decompressing it back to its original size. The goal of auto-encoder is to learn and extract the relevant information from the training data by minimizing the reconstruction error [29].

There are five layers in both the encoder and the decoder in five-layer auto-encoder module. The reason for using five layers is that after evaluating the effects of using two, three, four, five and six layers, it is determined that using five layers produces the best AUC and area under the precision–recall curve (AUPR) values. The encoder has five layers with neuron counts of 350, 250, 150, 100 and 64, while the decoder has five layers with neuron counts of 64, 100, 150, 250 and 350. We take the output of encoder layer as the characteristics of the disease.

Metabolite features extraction from GAE

GAE module is a tool used to get metabolite features, composed of an encoder and a decoder. The encoder has two layers, with 128 and 64 neurons, while the decoder also has two layers with 64 and 128 neurons. To enhance the ability of network to learn correlations between inputs, each layer is equipped with a self-attention mechanism [30].

The neural network takes in many different-sized vectors as input, and these vectors have some degree of interdependence among them. However, during the actual training, the relationship between these inputs cannot be given with full consideration, leading to poor training effect. Self-attention mechanism can address this limitation by allowing the network to identify connections among the different components of the input. This mechanism enables every input vector to output a new vector that incorporates the effects of all the other input vectors.

We obtain the attention matrix $\mathbf{ATTN}_0$ of the first encoder layer as follows:

$$\mathbf{ATTN}_0=\mathbf{SM}'\left((AW_0)\,V_{[0]}\right)+\mathbf{SM}'\left((AW_0)\,V_{[1]}\right)^T\tag{19}$$

A stands for metabolite and disease correlation matrix. W_0 is trainable weight matrix of the first encoder layer. V represents the trainable parameter and SM' denotes the integration similarity matrix of metabolites. At last, we get $\mathbf{ATTN}_0$ with a size of 2262*2262. Then, the attention matrix is normalized by Softmax.

Finally, the output H_0 of the first encoder layer is

$$H_0=\mathbf{ATTN}_0*(AW_0)\tag{20}$$

H_0 serves as the input of the second encoder layer. With the same method, we can calculate the attention value $\mathbf{ATTN}_1$ and output results H_1 for the second encoder layer as follows. W_1 is the weight matrix of the second encoder layer. Then, we take H_1 as the metabolite characteristics.

$$\mathbf{ATTN}_1=\mathbf{SM}'\left((H_0W_1)\,V_{[0]}\right)+\mathbf{SM}'\left((H_0W_1)\,V_{[1]}\right)^T\tag{21}$$

$$H_1=\mathbf{ATTN}_1*(H_0W_1)\tag{22}$$

When it comes to the decoder, it is worth noting that it shares its parameters with the encoder. For instance, the first decoder layer utilizes the same attention value $\mathbf{ATTN}_1$ with the second encoder layer, and also shares weight matrix W_1 with the second encoder layer. H_1 acts as the input of the first decoder layer. As a result, the output H_2 of the first decoder layer is as follows:

$$H_2=\mathbf{ATTN}_1*(H_1W_1)\tag{23}$$

Metabolite and disease features extraction from NMF

NMF, a matrix factorization method proposed by Lee and Seung *et al.* in 1999 [31], has been successfully applied in various fields. In previous studies, Ding *et al.* utilized NMF to extract the characteristics of miRNA and disease [28]. Here, we extend the application of NMF to obtain characteristics of metabolites and diseases. The core idea of NMF is to decompose a non-negative matrix $A_{\mathrm{m}*n}$ into two non-negative matrices, $U_{\mathrm{m}*k}$ and V_{k*n}, where $A_{m*n}\approx U_{m*k}V_{k*n}$, such that the product of $U_{\mathrm{m}*k}$ and V_{k*n} approximates the original matrix A_{m*n} as closely as possible. Specifically, U_{m*k} and V_{k*n} are referred to as the basis matrix and coefficient matrix, respectively.

Before and after factorization, the column vector of A is the weighted sum of all column vectors in U, and the weight coefficient is the element of the corresponding column vector of V. In general, m denotes the number of metabolites while n is the amount of diseases. k is smaller than m, and $(m+n)*k < mn$ is met. Tikhonov regularizationL_2 is used here to ensure the smoothness of U and V.

$$\mathbf{min}_{U\geq 0,V\geq 0}||W \odot (A-UV)\,||_F^2 + \lambda_1\|U\|_F^2 + \lambda_2\|V\|_F^2 \tag{24}$$

A is the adjacency matrix between metabolites and diseases, with a size of 2262*216. W takes the same value as A. $\lambda_1$and$\lambda_2$are the regularization coefficients, where $\lambda_1 = \lambda_2 = 0.01$ and $k=90$. $\odot$ represents Hadamard product, which means multiplying the corresponding elements of two matrices. $\|\cdot\|_F$ is to the matrix Frobenius norm.

NMF is a non-deterministic polynomial problem, which can be classified as an optimization problem and solved alternately by iterative methods. NMF algorithm provides a method of solving U and V based on simple iterations, which can help reduce the dimensions of high-dimensional data and is suitable for processing large-scale data.

Assuming that $\psi = (\varphi_{ik})$ and $\Phi = (\phi_{kj})$ are Lagrange multipliers, then the Lagrange function J of the optimization problem can be constructed as follows:

$$J(U,V) = \left||W \odot (A-UV)\,\right||_F^2 + \lambda_1 Tr\left(UU^T\right) + \lambda_2 Tr\left(VV^T\right) + Tr\left(\psi U^T\right) + Tr\left(\phi V^T\right) \tag{25}$$

Then, we calculate the partial derivatives of U and V:

$$\begin{aligned}\tfrac{\partial J}{\partial U} &= -2\left(W \odot (A-UV)\left(V^T\right)\right) + 2\lambda_1 U + \psi \\ &= -2\left((W \odot A)\,V^T\right) + 2\left(W \odot (UV)V^T\right) + 2\lambda_1 U + \psi\end{aligned} \tag{26}$$

$$\begin{aligned}\tfrac{\partial J}{\partial V} &= -2\left(U^T\,(W \odot (A-UV))\right) + 2\lambda_2 V + \phi \\ &= -2\left(U^T\,(W \odot A)\right) + 2\left(U^T\,(W \odot (UV))\right) + 2\lambda_2 V + \phi\end{aligned} \tag{27}$$

Due to the Karush–Kuhn–Tucker conditions $\varphi_{ik}u_{ik} = 0$ and $\phi_{kj}v_{kj} = 0$, the updating rule of U and V is further obtained:

$$\mathbf{u}_{ik}^{(t+1)} \leftarrow \mathbf{u}_{ik}^{(t)} \frac{\left((W \odot A)\,V^T\right)_{ik}}{\left(W \odot (UV)V^T + \lambda_1 U\right)_{ik}} \tag{28}$$

$$v_{kj}^{(t+1)} \leftarrow \mathbf{v}_{kj}^{(t)} \frac{\left(U^T\,(W \odot A)\right)_{kj}}{\left(U^T\,(W \odot (UV)) + \lambda_2 V\right)_{kj}} \tag{29}$$

U and V are updated 1000 times iteratively to obtain $U_{2262*90}$ and V_{90*216}. Here, 2262 represents the number of metabolites, while 216 denotes the number of diseases. $U_{2262*90}$ is used as the characteristic matrix of metabolites, while V_{216*90}^T is taken as the characteristic matrix of diseases.

Feature splicing

In summary, our approach leverages three modules to extract a total of 308 disease and metabolite features. A total of 64 disease features are extracted from five-layer auto-encoder module, 64 metabolite features are gained from GAE module, and 90 disease features and 90 metabolite features are obtained from NMF module. At last, we get $64+90+90+64=308$ features. Then, we splice the features obtained from these three modules into a long feature vector, so each of the metabolite–disease associations (MDAs) has a 308-dimensional feature vector. If the metabolite is associated with the disease in the dataset, the corresponding label is set to 1, otherwise 0. Finally, we send these feature vectors and corresponding labels to MLP training. Our approach thus integrates multiple modules and achieves a high-dimensional representation of MDAs, which enables more accurate classification.

MLP classifier

MLP, also named as artificial neural network [32], is a powerful tool used in data analysis. In addition to the input–output layer, it includes hidden layers, allowing it to make complex connections between data points. MDA-AENMF uses a fully connected MLP with three layers, each with a different number of neurons: 128, 64 and 1.

Since the ratio between positive and negative samples is imbalanced, we adopt 1:1 sampling strategy, selecting all positive samples and an equal number of randomly chosen negative samples. The features and corresponding labels of the training set are trained through MLP, and we take the results of the test set in MLP as the final predictive consequences.

RESULTS

Performance evaluation

To assess the performance of MDA-AENMF, we employ 5-fold CV. The dataset is divided into five equal parts, where each part is used once as the test set, while the remaining four parts are used as the training set. By repeating this process five times, we obtain a comprehensive evaluation of MDA-AENMF. We use some metrics to assess the efficacy of MDA-AENMF. The details of these metrics are defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{30}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{31}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{32}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{33}$$

In binary classification, a model may predict a sample as either positive or negative. TP represents a positive sample predicted to be positive, while FP is the negative sample predicted to be positive. Conversely, FN denotes the positive sample predicted to be negative, and TN is the negative sample predicted to be negative. Precision measures the probability that all the predicted positive samples are indeed positive, while Recall is the probability that a positive sample is predicted as positive.

The receiver operating characteristic (ROC) curve, also known as the susceptibility curve, is a graphical representation of binary classification performance. It plots FPR on the x-axis, which is the proportion of negative samples predicted to be positive in all negative samples, and TPR on the y-axis, which is the proportion of positive samples predicted to be positive in all positive samples. The AUC measures the classification effect, with a larger AUC indicating better performance. The AUPR is another measure of model performance, especially for unbalanced datasets. In the precision–recall diagram, the horizontal axis represents Recall,

Table 1. The main parameters of MDA-AENMF

Structure	Parameters	Value
KNNs for disease	k_1	21
KNNs for metabolite	k_2	226
Five-layer auto-encoder	Encoder/decoder layer number	5
	Neuron numbers of encoder layer	350, 250, 150, 100, 64
	Neuron numbers of decoder layer	64, 100, 150, 250, 350
	Optimizer	Adam
	Loss	Mse
	Epochs	20
	Batch_size	100
GAE	Encoder/decoder layer number	2
	Neuron numbers of encoder layer	128, 64
	Neuron numbers of decoder layer	64, 128
	Learning rate	Max: 0.005
NMF	λ_1	0.01
	λ_2	0.01
	k	90
MLP classifier	Layer number	3
	Neuron numbers of three layers	128, 64, 1
	Optimizer	Rmsprop
	Loss	Binary_crossentropy

and the vertical axis denotes Precision. A larger AUPR indicates that the model has better ability to correctly identify positive instances and to retrieve all positive instances.

We carefully select and set our parameters in Table 1 for our experiment. During the integration of disease similarity networks, we set k_1 to 21, which is the number of all diseases divided by 10. Similarly, during the integration of metabolite similarity networks, we set k_2 to 226, which is the number of all metabolites divided by 10. For five-layer auto-encoder module, we chose Adam as the optimizer and Mse as the loss function. In GAE module, we set the learning rate to a maximum of 0.005. For NMF module, we chose $\lambda_1 = \lambda_2 = 0.01$ and $k = 90$ as the regularization coefficients. For more detailed hyperparameters of MDA-AENMF, please refer to Table 1.

Comparison with other classifiers

To evaluate the effectiveness of MLP classifier, we conduct a comparison analysis with other popular classifiers including Bagging, Random Forest, AdaBoost, Gradient boosting (GBDT), XGBoost and LightGBM. The results of this analysis are presented in Table 2. Impressively, MLP classifier outperforms all other classifiers with the highest AUC and AUPR scores of 0.975 and 0.973, respectively. This indicates that MLP can better distinguish between positive and negative samples and has higher precision for positive samples. In contrast, Bagging classifier comes in second place with AUC of 0.941 and AUPR of 0.955. In comparison, AdaBoost classifier shows the poorest performance, with the lowest AUC and AUPR scores of 0.904 and 0.928, respectively. The best results of MLP on AUC and AUPR may be attributed to the application of neural networks to train features, while other classifiers employ machine learning ensemble classification algorithms to train features. MLP classifier can deal with various complex nonlinear problems, can adaptively learn features, and has robustness to a certain degree of noise and outliers. However, Bagging classifier requires training multiple base classifiers, which may lead to bias in model prediction due to the possibility of some samples not being sampled in the random subset used to train base classifiers. Random Forest classifier requires training multiple decision trees, which may lead to overfitting. In addition, the performance of Random Forest classifier may be affected if there is noise or outliers in the data. GBDT classifier is prone to overfitting because each tree is built on the basis of the residual of the previous tree. In addition, it is susceptible to the influence of noise and outliers, which can result in poor performance. Taking the above into consideration, we adopt MLP classifier to improve predictive performance of our model.

Table 2. Comparison analysis between MLP classifier and other popular classifiers on the same dataset

Classifier	AUPR	AUC
MLP	**0.973**	**0.975**
Bagging	0.955	0.941
RandomForest	0.952	0.937
AdaBoost	0.928	0.904
GBDT	0.952	0.936
XGBoost	0.931	0.919
LightGBM	0.948	0.931

The bold values are the AUC and AUPR scores when MLP classifier is adopted in MDAAENMF.

Comparison with previous methods

To evaluate the performance of MDA-AENMF, we conduct comparative experiments with five state-of-the-art binary classification models in the field of bioinformatics, namely, RWR [19], PageRank [33], KATZ [20], EKRR [34] and GCNAT [25]. In order to enhance the persuasiveness of the comparative experiments, our comparative models cover network-based methods, machine learning-based methods and deep learning-based methods.

(i) RWR calculates pairwise metabolite similarity based on disease sets, constructs a weighted metabolite association network and uses random walk to predict novel metabolic markers of diseases.

(ii) PageRank is an algorithm based on graph theory. Each node in the network has a PageRank value, indicating the importance of the node.

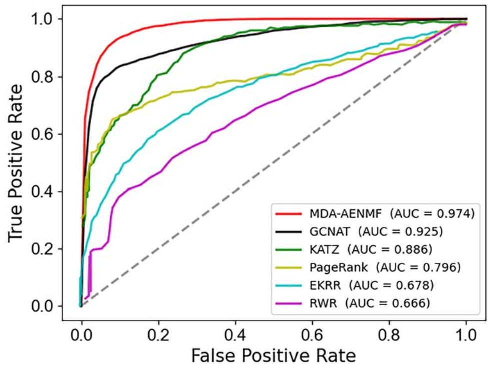

Figure 2. ROC curves of MDA-AENMF and comparison methods by 5-fold CV under the same dataset.

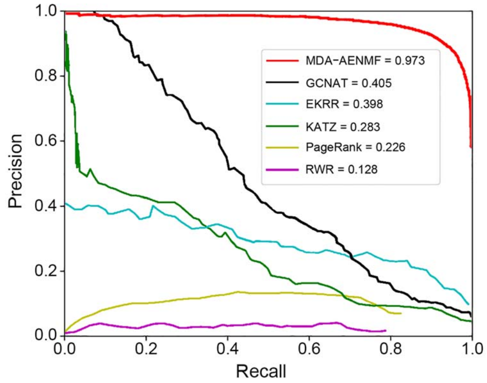

Figure 3. PR curves of MDA-AENMF and comparison methods by 5-fold CV under the same dataset.

(iii) KATZ is the first model to apply the KATZ algorithm to the field of metabolomics. The KATZ index can distinguish the different influence of distinct neighbor nodes.

(iv) EKRR obtains features by integrating multiple similarity networks. Multiple base classifiers are obtained by combining two Kernel Ridge Regression classifiers from the miRNA and disease sides, respectively, based on random selection of features. Then, average strategy for these base classifiers is adopted to obtain final association scores.

(v) GCNAT is an algorithm based on graph neural network, and each layer of the network is added attention mechanism to better extract features.

To ensure the authenticity and accuracy of the comparison, we evaluate all six models on the same dataset. As illustrated in Figure 2, MDA-AENMF achieves the highest AUC value of 0.974, surpassing the other models by a significant margin. Specifically, MDA-AENMF outperforms RWR, EKRR, PageRank, KATZ and GCNAT by 30.8%, 29.6%, 17.8%, 8.8% and 4.9% in terms of AUC, respectively. The observed discrepancy between PageRank and MDA-AENMF may stem from the former lack of consideration for biological similarity, which results in lower evaluation scores. Although other models incorporate similarity measures, their similarity networks are fewer than those used by MDA-AENMF, which employs a nonlinear integration method. In contrast, the other models, except for GCNAT, rely on optimization and complex network algorithms, without incorporating neural networks and attention mechanisms, thereby resulting in lower performance. As depicted in Figure 3, MDA-AENMF demonstrates a remarkable improvement compared to other algorithms in terms of AUPR, exhibiting 84.5%, 74.7%, 69%, 57.5% and 56.8% higher performance than that of RWR, PageRank, KATZ, EKRR and GCNAT, respectively. This impressive performance can be attributed to our adoption of a 1:1 sampling strategy for positive and negative samples. By comparing MDA-AENMF to these state-of-the-art models, we can demonstrate its superior performance in binary classification tasks in the field of bioinformatics.

External validation

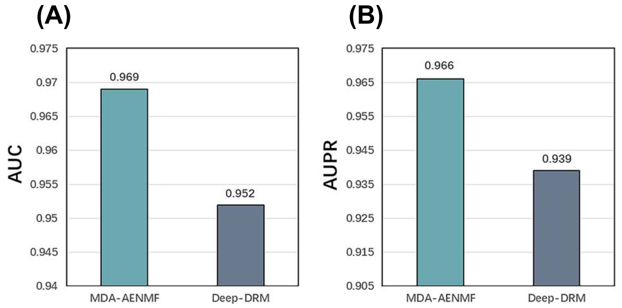

Figure 4. Comparison results between MDA-AENMF and Deep-DRM.

To further demonstrate the impressive generalization ability of MDA-AENMF, we conduct a comparison by using the same dataset with Deep-DRM [22]. As shown in Figure 4, the AUC and AUPR achieved by Deep-DRM are 0.952 and 0.939, respectively. However, the AUC and AUPR obtained by MDA-AENMF are 1.7% and 2.7% higher than those of Deep-DRM, indicating that MDA-AENMF outperforms Deep-DRM to a certain extent. This superiority may be attributed to the utilization of multiple metabolite and disease similarity networks and incorporation of attention mechanisms during feature extraction.

Case studies

To further analyze the performance of MDA-AENMF and the accuracy of predictive results, we conduct case studies on four diseases, namely, leukemia, uremia, obesity and hepatitis. Our objective is to identify associations present in the prediction results that are not present in the original dataset and then validate the top 20 ranked new associations. Leukemia is a malignant tumor disease of hematopoietic stem cells in the hematopoietic system [35]. There are four main symptoms of leukemia including fever, bleeding in gums and nose, bone pain and anemia [36]. Through a case study of leukemia, we examine the top 20 new associations and the validated results are shown in Table 3. For instance, hypoxanthine and its nucleoside enhance differentiation induction properties of HL-60 promyelocytic leukemia cells [37]. Choline-magnesium trisalicylate regulates gene expression in acute myeloid leukemia during induction chemotherapy [38]. In addition, the apoptotic effect of lactic acid bacteria on human T leukemia cell line is related to the activity of arginine deiminase and/or sphingomyelin enzyme [39].

Uremia is a chronic renal failure end-stage syndrome that can cause a range of distressing symptoms, including nausea, vomiting, edema, bleeding and arrhythmia [40]. Through case analysis of uremia, 18 of 20 new associations have been verified as shown in Table 4. For example, in patients with chronic renal failure, intracellular concentrations of valine, threonine, lysine and carnosine are lower. In addition, serine, tyrosine and taurine are also usually low on low protein diets and hemodialysis [41]. There is an increased production and release of alanine and glutamine in skeletal muscle of rats with chronic uremia, and this increase appears to be due in part to enhanced net protein degradation [42].

Table 3. Top 20 potential metabolites associated with leukemia

Leukemia

Rank	Metabolite name	Evidences	Confirmed	PMID
1	Citrulline	HMDB0000904	Yes	32115690
2	Hypoxanthine	HMDB0000157	Yes	3855287
3	L-Aspartic acid	HMDB0000191	Yes	22356135
4	Choline	HMDB0000097	Yes	27659510
5	Creatinine	HMDB0000562	Yes	13684930
6	Creatine	HMDB0000064	Yes	13684930
7	*p*-Hydroxyphenylacetic acid	HMDB0000020	No	–
8	L-Lactic acid	HMDB0000190	Yes	11962255
9	Betaine	HMDB0000043	Yes	14950203
10	Trimethylamine *N*-oxide	HMDB0000925	No	–
11	Citric acid	HMDB0000094	No	–
12	Homovanillic acid	HMDB0000118	No	–
13	Pyruvic acid	HMDB0000243	Yes	13778379
14	L-Proline	HMDB0000162	Yes	29307398
15	Homocysteine	HMDB0000742	Yes	1988122
16	Hippuric acid	HMDB0000714	No	–
17	gamma-Aminobutyric acid	HMDB0000112	No	–
18	Succinic acid	HMDB0000254	Yes	33199038
19	Testosterone	HMDB0000234	Yes	9783810
20	Bilirubin	HMDB0000054	Yes	35658244

Table 4. Top 20 potential metabolites associated with uremia

Uremia

Rank	Metabolite name	Evidences	Confirmed	PMID
1	L-Tyrosine	HMDB0000158	Yes	237643
2	Taurine	HMDB0000251	Yes	2674259
3	L-Arginine	HMDB0000517	Yes	12883450
4	Homovanillic acid	HMDB0000118	No	–
5	L-Lysine	HMDB0000182	Yes	2674258
6	L-Glutamine	HMDB0000641	Yes	690188
7	L-Phenylalanine	HMDB0000159	Yes	2674258
8	L-Serine	HMDB0000187	Yes	2674259
9	Glycine	HMDB0000123	Yes	237643
10	L-Tryptophan	HMDB0000929	Yes	11867954
11	Vanylglycol	HMDB0001490	No	–
12	L-Histidine	HMDB0000177	Yes	4753036
13	Citric acid	HMDB0000094	Yes	13475172
14	L-Isoleucine	HMDB0000172	Yes	2674258
15	L-Valine	HMDB0000883	Yes	685880
16	Creatinine	HMDB0000562	Yes	14799499
17	L-Alanine	HMDB0000161	Yes	690188
18	Citrulline	HMDB0000904	Yes	2079537
19	Betaine	HMDB0000043	Yes	7301006
20	L-Threonine	HMDB0000167	Yes	2674258

Obesity is a chronic metabolic disease that is characterized by the accumulation of excess weight and fat in the body [43]. Severe obesity can have serious consequences for overall health, leading to abnormal metabolism and placing a burden on various organs. We investigate the top 20 new associations through case studies of obesity, and the results are presented in Table 5. For example, obese women typically have higher average serum creatinine and creatinine clearance rates than normal healthy women [44]. In addition, research has shown that dietary supplementation with L-leucine and L-alanine can have an acute effect in preventing obesity caused by a high-fat diet [45].

Hepatitis is an inflammation caused by damage to liver cells. Clinical symptoms include fatigue, bloating, anorexia, emaciation, yellowish face, and even nausea and vomiting [46]. Through case analysis of hepatitis, we examine the top 20 new associations, 15 of which have been validated in Table 6. For example, research has shown that calcium phosphate nanoparticles can provide some immunity against hepatitis B virus genes *in vitro* and *in vivo* [47]. Furthermore, in more severe cases of cholestasis or liver

Table 5. Top 20 potential metabolites associated with obesity

Obesity				
Rank	**Metabolite name**	**Evidences**	**Confirmed**	**PMID**
1	Cholesterol	HMDB0000067	Yes	20025694
2	Creatinine	HMDB0000562	Yes	34605468
3	Dopamine	HMDB0000073	Yes	26514168
4	5-Hydroxyindoleacetic acid	HMDB0000763	Yes	6184736
5	Cortisol	HMDB0000063	Yes	27345309
6	Glycerol	HMDB0000131	Yes	1607071
7	L-Alanine	HMDB0000161	Yes	22847780
8	Iron	HMDB0015531	Yes	35466128
9	Trimethylamine *N*-oxide	HMDB0000925	Yes	32017391
10	1-Methyladenosine	HMDB0003331	No	–
11	L-Lysine	HMDB0000182	No	–
12	Androstenedione	HMDB0000053	No	–
13	Creatine	HMDB0000064	Yes	28885625
14	L-Lactic acid	HMDB0000190	Yes	24232731
15	Epinephrine	HMDB0000068	Yes	11229419
16	Testosterone	HMDB0000234	Yes	25982085
17	D-Glucose	HMDB0000122	Yes	5040914
18	Homovanillic acid	HMDB0000118	No	–
19	24-Hydroxycholesterol	HMDB0001419	No	–
20	Uracil	HMDB0000300	No	–

dysfunction, the excretion of cobioline I in the biliary tract is reduced [48]. Finally, neuropeptide Y and substance P released by nerve fibers and immune cells are believed to play a role in inflammation and elimination of inflammation in hepatitis [49].

In addition to the above four diseases, we have also done case studies of other nine diseases, and we have selected the top 20 new associations to verify for each of those diseases. For details, please refer to the Supplementary Materials.

Ablation experiments

To further evaluate the generalization and robustness of MDA-AENMF, we conduct a series of ablation experiments on the three feature extraction modules to confirm the contribution of each module. In the ablation experiments, we remove one module at a time while keeping the other two modules unchanged. To ensure the accuracy of our results, we perform these experiments on the same data samples.

(i) Del-DAE: five-layer auto-encoder is deleted from MDA-AENMF.
(ii) Del-NMF: the application of NMF is abandoned.
(iii) Del-MAE: GAE is no longer in use.

As shown in Figure 5, we can see that MDA-AENMF achieves the highest AUC and AUPR values of 0.975 and 0.973, respectively. The AUC and AUPR of Del-DAE are 0.947 and 0.945, respectively. Compared with MDA-AENMF, AUC decreases by 2.8% and AUPR reduces by 2.8%, indicating that the disease characteristics extracted by the five-layer auto-encoder are significant for our training. When we perform Del-NMF, it produces inferior results, with AUC and AUPR values of 0.834 and 0.806, respectively. This underscores the importance of both the 90 disease features and 90 metabolite features obtained by NMF, which are critical components of MDA-AENMF. Finally, when we train Del-MAE, we could see that the value of each index is slightly lower than that of MDA-AENMF, revealing that GAE also plays its role in our characteristic training. To sum up, one of these three modules is indispensable, and all features are crucial for predicting the potential associations between metabolites and diseases.

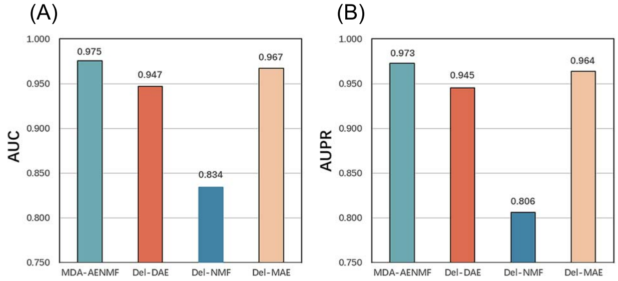

Figure 5. Comparison analysis between MDA-AENMF and its ablation experiments.

DISCUSSION AND CONCLUSION

With the rapid advancement of medical technology, it is increasingly apparent that the relationship between diseases and metabolites is becoming closer. The occurrence of each disease will be accompanied by specific changes in the level of some metabolites in the human body. Accurately identifying and quantifying these changes can help in assessing the severity of the disease and determining appropriate treatment options. However, the traditional research methods can be time consuming and expensive. In response to this challenge, researchers have started to explore the application of optimization algorithms and complex network algorithms to predict the relationship between metabolites and diseases. In past studies, Hu *et al.* used random walk to identify disease-related metabolites in 2018 [19]. Weighted metabolite association network is constructed by using the similarity of all metabolite pairs. This network is

Table 6. Top 20 potential metabolites associated with hepatitis

Hepatitis				
Rank	**Metabolite name**	**Evidences**	**Confirmed**	**PMID**
1	Cholesterol	HMDB0000067	Yes	35409259
2	D-Glucose	HMDB0000122	Yes	23347806
3	Homocysteine	HMDB0000742	Yes	17483780
4	Phosphate	HMDB0001429	Yes	32344172
5	Dopamine	HMDB0000073	Yes	–
6	1-Methyladenosine	HMDB0003331	No	–
7	Coproporphyrin I	HMDB0000643	Yes	10735544
8	Quinolinic acid	HMDB0000232	Yes	–
9	Bilirubin	HMDB0000054	Yes	33516950
10	Substance P	HMDB0001897	Yes	26568102
11	Cortisol	HMDB0000063	Yes	20848846
12	Androstenedione	HMDB0000053	Yes	25912488
13	Heptacarboxylporphyrin I	HMDB0000737	No	–
14	Glycine	HMDB0000123	Yes	31966360
15	24-Hydroxycholesterol	HMDB0001419	No	–
16	Uric acid	HMDB0000289	Yes	33978375
17	Leukotriene C4	HMDB0001198	Yes	10770113
18	Calcium	HMDB0000464	Yes	29307794
19	Coproporphyrin III	HMDB0000570	Yes	29856826
20	Iron	HMDB0015531	Yes	24251712

then used to predict new metabolic markers of disease using random walks. Later, Lei *et al.* developed a calculation method based on KATZ to predict the metabolite–disease association [20]. More reliable disease similarity is obtained by integrating disease semantic similarity and improved disease Gaussian Interaction profile kernel similarity. In 2020, Lei *et al.* proposed a linear neighborhood similarity with improved bipartite network projection algorithm [21]. The linear neighborhood similarity matrix of metabolites (diseases) is reconstructed based on the new characteristics obtained from the known metabolite–disease associations and the combined similarity. Then, an improved bipartite network projection method is used to predict associations between metabolites and diseases. Furthermore, Zhao *et al.* introduced a Deep-DRM model in 2021 [22], which codes metabolites and disease similarity networks, respectively, by using graph convolution network to obtain corresponding characteristics. Then, PCA is used to reduce the dimensions of features. Finally, based on these features, a deep neural network is constructed to identify MDAs. In the same year, Zhang *et al.* proposed a calculation method called LGBMMDA [23]. This method extracts features from statistical measurements, graph theory measurements and matrix factorization results, uses PCA to remove noise or redundancy, and finally sends these features to the classifier LGBM for prediction. In 2022, Tie *et al.* proposed a metabolite–disease association prediction algorithm based on DeepWalk and random forest [24]. DeepWalk is used to extract metabolite characteristics based on metabolite–gene association network, and random forest is used to predict the association between metabolite and disease. During the same period, Sun *et al.* came up with a graph neural network to predict the associations between metabolites and diseases [25]. A heterogeneous network is constructed based on the metabolite–disease associations, metabolite similarities and disease similarities. Each layer of graph neural network is added attention mechanism to precisely concentrate on crucial information within the input data. While these methods have shown some promise, there have been few instances where neural networks and attention mechanisms are simultaneously employed. The number of established networks for metabolite and disease similarities is also limited, and the integration of these networks often relies on linear methods, which fail to effectively combine various features. In addition, there is room for further improvement in the accuracy of their predictions. To address those issues, we propose a novel deep learning method named MDA-AENMF to predict the potential correlations between metabolites and diseases. This method integrates multiple similarities for diseases and metabolites, and employs a variety of feature extraction techniques, including five-layer auto-encoder for disease features, NMF for metabolite and disease features, and GAE for metabolite features. These features are then combined into a long vector and fed into MLP classifier for training. Our method has achieved a remarkable AUC of 0.975 and an AUPR of 0.973 by 5-fold CV, demonstrating its superiority over previous methods. We also conduct case studies on four diseases, including leukemia, uremia, obesity and hepatitis, which further confirm the effectiveness and predictive power of MDA-AENMF.

The good performance of MDA-AENMF can be attributed to several key factors. First, we compute DSS, DGIP and DSIE for diseases, and MSS, MGIP and MSIE for metabolites. The computation of these similarity networks greatly increases biological information and leads to more illustrative predictive results. Second, the application of nonlinear integration method to combine these similarity networks improves the accuracy of similarity measurements compared with traditional linear integration methods. Third, we employ three feature extraction modules for both metabolites and diseases, which are indispensable to achieving our excellent predictive results. At last, we combine all of the extracted features for each metabolite–disease pair to form a more comprehensive vector of high-quality feature information, and then send this feature vector and corresponding tag to the classifier for training. The impressive performance of MDA-AENMF also confirms that the features we extract are meaningful and high-quality..

However, MDA-AENMF still has a few shortcomings. First, the obtained metabolite and disease association matrix is a sparse matrix, with a significant imbalance between positive and negative samples, where positive samples are outnumbered by negative samples by a ratio of almost 1:100. Second, although we have calculated a variety of similar networks for metabolites or diseases, the increased biological information results in noise, and some features extracted from multiple modules are redundant. Third, further optimization of certain parameters may be necessary to achieve better results. How to overcome these problems is also a challenge in our future study. Nevertheless, we are confident that with the confirmation of more metabolite–disease associations and the development of more efficient parameter optimization algorithms, we can enhance the predictive performance of MDA-AENMF.

Key Points

- We calculate multiple similarity networks for diseases and metabolites, and integrate them in a nonlinear way.
- We leverage advanced techniques such as five-layer auto-encoder, NMF and GAE to extract dimension reduction features for metabolites and diseases.
- We construct a long feature vector for each metabolite–disease pair by splicing together their respective features. These feature vectors, along with their corresponding tags, are used to train MLP classifier.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2021ZD0201900), National Natural Science Foundation of China (Grant Nos. 12090052 and 11805091), Foundation of Education Department of Liaoning Province (Grant Nos. LJKZ0280 and LJKZ0320) and Natural Science Foundation of Liaoning Province (Grant No. 2023-MS-288).

DATA AVAILABILITY

The codes and datasets are available online at https://github.com/zhaoqi106/MDA-AENMF.

REFERENCES

1. Eckel RH, Grundy SM, Zimmet PZ. The metabolic syndrome. *Lancet* 2005;**365**:1415–28. https://doi.org/10.1016/S0140-6736(05)66378-7.
2. Kolotkin RL, Meter K, Williams GR. Quality of life and obesity. *Obes Rev* 2001;**2**:219–29. https://doi.org/10.1046/j.1467-789x.2001.00040.x.
3. Powell EE, Wong VW, Rinella M. Non-alcoholic fatty liver disease. *Lancet* 2021;**397**:2212–24. https://doi.org/10.1016/S0140-6736(20)32511-3.
4. Leonard EA, Marshall RJ. Cardiovascular disease in women. *Prim Care* 2018;**45**:131–41. https://doi.org/10.1016/j.pop.2017.10.004.
5. Wu J, Wang K, Wang X, *et al.* The role of the gut microbiome and its metabolites in metabolic diseases. *Protein Cell* 2021;**12**:360–73. https://doi.org/10.1007/s13238-020-00814-7.
6. Taylor K, McBride N, Zhao J, *et al.* The relationship of maternal gestational mass spectrometry-derived metabolites with offspring congenital heart disease: results from multivariable and Mendelian randomization analyses. *J Cardiovasc Dev Dis* 2022;**9**:237. https://doi.org/10.3390/jcdd9080237.
7. Azam S, Park JY, Kim IS, Choi DK. Piperine and its metabolite's pharmacology in neurodegenerative and neurological diseases. *Biomedicine* 2022;**10**:154. https://doi.org/10.3390/biomedicines10010154.
8. Lord J, Green R, Choi SW, *et al.* Disentangling independent and mediated causal relationships between blood metabolites, cognitive factors, and Alzheimer's disease. *Biol Psychiatry Glob Open Sci* 2022;**2**:167–79. https://doi.org/10.1016/j.bpsgos.2021.07.010.
9. Wang T, Sun J, Zhao Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. *Comput Biol Med* 2023;**153**:106464. https://doi.org/10.1016/j.compbiomed.2022.106464.
10. Liu H, Ren G, Chen H, *et al.* Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl Based Syst* 2020;**191**:105261. https://doi.org/10.1016/j.knosys.2019.105261.
11. Zhang L, Yang P, Feng H, *et al.* Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscip Sci Comput Life Sci* 2021;**13**:535–45. https://doi.org/10.1007/s12539-021-00458-z.
12. Wang W, Zhang L, Sun J, *et al.* Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field. *Brief Bioinform* 2022;**23**:bbac463. https://doi.org/10.1093/bib/bbac463.
13. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: taxonomy, trends and challenges of computational models. *Brief Bioinform* 2022;**23**:bbac358. https://doi.org/10.1093/bib/bbac358.
14. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: towards systematic evaluation of computational models. *Brief Bioinform* 2022;**23**:bbac407. https://doi.org/10.1093/bib/bbac407.
15. Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2019;**20**:515–39. https://doi.org/10.1093/bib/bbx130.
16. Wang CC, Han CD, Zhao Q, Chen X. Circular RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2021;**22**:bbab286. https://doi.org/10.1093/bib/bbab286.
17. Zhao Q, Yang Y, Ren G, *et al.* Integrating bipartite network projection and KATZ measure to identify novel circRNA-disease associations. *IEEE Trans Nanobioscience* 2019;**18**:578–84. https://doi.org/10.1109/tnb.2019.2922214.
18. Ge E, Yang Y, Gang M, *et al.* Predicting human disease-associated circRNAs based on locality-constrained linear coding. *Genomics* 2020;**112**:1335–42. https://doi.org/10.1016/j.ygeno.2019.08.001.
19. Hu Y, Zhao T, Zhang N, *et al.* Identifying diseases-related metabolites using random walk. *BMC Bioinform* 2018;**19**:116. https://doi.org/10.1186/s12859-018-2098-1.
20. Lei X, Zhang C. Predicting metabolite-disease associations based on KATZ model. *BioData Min* 2019;**12**:19. https://doi.org/10.1186/s13040-019-0206-z.
21. Lei X, Zhang C. Predicting metabolite-disease associations based on linear neighborhood similarity with improved bipartite network projection algorithm. *Complexity* 2020;**2020**:1–11. https://doi.org/10.1155/2020/9342640.

22. Zhao T, Hu Y, Cheng L. Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief Bioinform* 2021;**22**:bbaa212. https://doi.org/10.1093/bib/bbaa212.
23. Zhang C, Lei X, Liu L. Predicting metabolite-disease associations based on LightGBM model. *Front Genet* 2021;**12**:660275. https://doi.org/10.3389/fgene.2021.660275.
24. Tie J, Lei X, Pan Y. Metabolite-disease association prediction algorithm combining DeepWalk and random forest. *Tsinghua Sci Technol* 2022;**27**:58–67. https://doi.org/10.26599/tst.2021.9010003.
25. Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief Bioinform* 2022;**23**:bbac266. https://doi.org/10.1093/bib/bbac266.
26. Fang Z, Lei X. Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network. *Big Data Min Anal* 2019;**2**:261–72. https://doi.org/10.26599/bdma.2019.9020010.
27. Li X, Lin Y, Gu C, Yang J. FCMDAP: using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs. *BMC Syst Biol* 2019;**13**:26. https://doi.org/10.1186/s12918-019-0696-9.
28. Ding Y, Lei X, Liao B, Wu FX. Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization. *IEEE J Biomed Health Inform* 2022;**26**:446–57. https://doi.org/10.1109/JBHI.2021.3088342.
29. Liu W, Lin H, Huang L, *et al.* Identification of miRNA-disease associations via deep forest ensemble learning based on autoencoder. *Brief Bioinform* 2022;**23**:bbac104. https://doi.org/10.1093/bib/bbac104.
30. Deng L, Liu Z, Qian Y, Zhang J. Predicting circRNA-drug sensitivity associations via graph attention auto-encoder. *BMC Bioinform* 2022;**23**:160. https://doi.org/10.1186/s12859-022-04694-y.
31. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**:788–91. https://doi.org/10.1038/44565.
32. Riedmiller M, Lernen A. Multi layer perceptron. *Machine Learning Lab Special Lecture*, University of Freiburg, 2014:7–24.
33. Yates EJ, Dixon LC. PageRank as a method to rank biomedical literature by importance. *Source Code Biol Med* 2015;**10**:16. https://doi.org/10.1186/s13029-015-0046-2.
34. Peng LH, Zhou LQ, Chen X, Piao X. A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front Bioeng Biotechnol* 2020;**8**:40. https://doi.org/10.3389/fbioe.2020.00040.
35. Dzierzak E, Bigas A. Blood development: hematopoietic stem cell dependence and independence. *Cell Stem Cell* 2018;**22**:639–51. https://doi.org/10.1016/j.stem.2018.04.015.
36. Leak Bryant A, Lee Walton A, Shaw-Kokot J, *et al.* Patient-reported symptoms and quality of life in adults with acute leukemia: a systematic review. *Oncol Nurs Forum* 2015;**42**:E91–e101. https://doi.org/10.1188/15.Onf.E91-e101.
37. Ishiguro K, Sartorelli AC. Enhancement of the differentiation-inducing properties of 6-thioguanine by hypoxanthine and its nucleosides in HL-60 promyelocytic leukemia cells. *Cancer Res* 1985;**45**:91–5.
38. Medina D, David K, Lin Y, *et al.* Choline-magnesium trisalicylate modulates acute myelogenous leukemia gene expression during induction chemotherapy. *Leuk Lymphoma* 2017;**58**:1227–30. https://doi.org/10.1080/10428194.2016.1225206.
39. Di Marzio L, Russo FP, D'Alò S, *et al.* Apoptotic effects of selected strains of lactic acid bacteria on a human T leukemia cell line are associated with bacterial arginine deiminase and/or sphingomyelinase activities. *Nutr Cancer* 2001;**40**:185–96. https://doi.org/10.1207/s15327914nc402_16.
40. Haller H, Strauer BE. Renal failure. *Internist (Berl)* 2012;**53**:789–90. https://doi.org/10.1007/s00108-011-2980-7.
41. Rudman D, Rudman IW, Mattson DE, *et al.* Fractures in the men of a veterans administration nursing home: relation to 1,25-dihydroxyvitamin D. *J Am Coll Nutr* 1989;**8**:324–34. https://doi.org/10.1080/07315724.1989.10720308.
42. Garber AJ. Skeletal muscle protein and amino acid metabolism in experimental chronic uremia in the rat: accelerated alanine and glutamine formation and release. *J Clin Invest* 1978;**62**:623–32. https://doi.org/10.1172/jci109169.
43. Xu S, Xue Y. Pediatric obesity: causes, symptoms, prevention and treatment. *Exp Ther Med* 2016;**11**:15–20. https://doi.org/10.3892/etm.2015.2853.
44. Meherubin I, Nessa A, Huda MN, *et al.* Level of serum creatinine and creatinine clearance rate in obese female. *Mymensingh Med J* 2021;**30**:991–6.
45. Freudenberg A, Petzke KJ, Klaus S. Dietary L-leucine and L-alanine supplementation have similar acute effects in the prevention of high-fat diet-induced obesity. *Amino Acids* 2013;**44**:519–28. https://doi.org/10.1007/s00726-012-1363-2.
46. Turner GC. Hepatitis. *Br Med J* 1973;**1**:476–9. https://doi.org/10.1136/bmj.1.5851.476.
47. Rojas-Sánchez L, Zhang E, Sokolova V, *et al.* Genetic immunization against hepatitis B virus with calcium phosphate nanoparticles in vitro and in vivo. *Acta Biomater* 2020;**110**:254–65. https://doi.org/10.1016/j.actbio.2020.04.021.
48. Gibson PR, Grant J, Cronin V, *et al.* Effect of hepatobiliary disease, chronic hepatitis C and hepatitis B virus infections and interferon-alpha on porphyrin profiles in plasma, urine and faeces. *J Gastroenterol Hepatol* 2000;**15**:192–201. https://doi.org/10.1046/j.1440-1746.2000.02065.x.
49. Fehér E. Changes in neuropeptide Y and substance P immunoreactive nerve fibres and immunocompetent cells in hepatitis. *Orv Hetil* 2015;**156**:1892–7. https://doi.org/10.1556/650.2015.30300.