frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 05 June 2023
DOI 10.3389/fpsyt.2023.1149947

OPEN ACCESS

EDITED BY
Tao Wang,
Northwestern Polytechnical University, China

REVIEWED BY
Chenchen Li,
University of Pennsylvania, United States
Xueqing Chen,
Brigham and Women's Hospital and Harvard Medical School, United States

*CORRESPONDENCE
Yuanyuan Ma
✉ chonghua_1983@126.com

RECEIVED 10 February 2023
ACCEPTED 10 May 2023
PUBLISHED 05 June 2023

CITATION
Zhao Y, Ma Y and Zhang Q (2023) Metabolite-disease interaction prediction based on logistic matrix factorization and local neighborhood constraints.
Front. Psychiatry 14:1149947.
doi: 10.3389/fpsyt.2023.1149947

Metabolite-disease interaction prediction based on logistic matrix factorization and local neighborhood constraints

Yongbiao Zhao[1,2], Yuanyuan Ma[2]* and Qilin Zhang[2]

[1]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, Hubei, China, [2]School of Computer Engineering, Hubei University of Arts and Science, Xiangyang, Hubei, China

Background: Increasing evidence indicates that metabolites are closely related to human diseases. Identifying disease-related metabolites is especially important for the diagnosis and treatment of disease. Previous works have mainly focused on the global topological information of metabolite and disease similarity networks. However, the local tiny structure of metabolites and diseases may have been ignored, leading to insufficiency and inaccuracy in the latent metabolite-disease interaction mining.

Methods: To solve the aforementioned problem, we propose a novel metabolite-disease interaction prediction method with logical matrix factorization and local nearest neighbor constraints (LMFLNC). First, the algorithm constructs metabolite-metabolite and disease-disease similarity networks by integrating multi-source heterogeneous microbiome data. Then, the local spectral matrices based on these two networks are established and used as the input of the model, together with the known metabolite-disease interaction network. Finally, the probability of metabolite-disease interaction is calculated according to the learned latent representations of metabolites and diseases.

Results: Extensive experiments on the metabolite-disease interaction data were conducted. The results show that the proposed LMFLNC method outperformed the second-best algorithm by 5.28 and 5.61% in the AUPR and F1, respectively. The LMFLNC method also exhibited several potential metabolite-disease interactions, such as "Cortisol" (HMDB0000063), relating to "21-Hydroxylase deficiency," and "3-Hydroxybutyric acid" (HMDB0000011) and "Acetoacetic acid" (HMDB0000060), both relating to "3-Hydroxy-3-methylglutaryl-CoA lyase deficiency."

Conclusion: The proposed LMFLNC method can well preserve the geometrical structure of original data and can thus effectively predict the underlying associations between metabolites and diseases. The experimental results show its effectiveness in metabolite-disease interaction prediction.

KEYWORDS

logistic matrix factorization, neighborhood regularization, metabolite-disease interaction, association prediction, vicus matrix

1. Introduction

Metabolites, the final product of the cell regulation process, are also regarded as the final response of a biological system to genetic or environmental changes (1, 2). Changes in metabolite levels are important markers of disease development, directly reflecting the physiological state of the human body and metabolic abnormalities. Nicholson et al. (3) pointed out that the level of metabolites reflects the effect of the human body on drug treatment and can be used as an important indicator of susceptibility and disease rehabilitation. Disease-related metabolite identification can improve clinical diagnosis and deepen the understanding of pathological mechanisms. Therefore, it is a critical task and challenge in precision medicine and biology (4).

Researchers have developed numerous methods, mostly experimental or computational, to mine the relationship between metabolites and diseases. For example, Ouyang et al. (5) discovered that metabolites (e.g., isoleucine, triglyceride, leucine, and creatinine) revealed significantly higher in the serum of pancreatic cancer patients than those in the serum of healthy controls by using 1H NMR spectroscopy and principal component analysis. Reinke et al. (6) did a metabolomics analysis to identify different metabotypes of asthma severity and found that 15 out of 66 identified serum metabolites were significantly changed with asthma. Ibanez et al. (7) developed a non-targeted metabolomics method to detect differences in metabolites in cerebrospinal fluid samples from subjects with different cognitive states associated with the progression of Alzheimer's disease. Further, Wang et al. (8) proposed a metabolomics method based on ultra-high performance liquid chromatography–mass spectrometry to identify 13 potential biomarkers, such as succinic acid (Canavaninosuccinate) and glycochenodeoxycholic acid, which effectively distinguished patients with hepatocellular carcinoma or cirrhosis from the control group and provided important indicators for the early diagnosis and screening of patients with liver cancer. Compared with traditional experimental methods, computational approaches are relatively convenient and economical and are now more important in the field of disease-metabolite interaction relationship prediction.

Recently, some researchers have used machine learning methods to predict the interactions between metabolites and diseases (1, 2, 9–12). The majority of these methods work as follows: First, a metabolite-related heterogeneous network is built by integrating multi-omics information; second, the candidate metabolites are scored via a random walk-based method (4, 9, 13); finally, the ranking of disease-related metabolites is obtained according to the score. These methods comprehensively consider the information from multiple sources, including the genome, phenotype, and metabolic pathway, but they ignore the noise and outliers in the metabolite interaction network, undermining the reliability of the final prediction. An effective solution is to utilize the neighbor information of disease (metabolite) nodes. It benefits in two aspects: (i) effectively reducing the computational complexity, especially the construction of large-scale node similarity networks, and (ii) largely eliminating noise and interference information.

Several studies have verified that compared with the global similarity network, the local structure information (neighbors) of nodes can significantly improve the algorithm's performance. Ma et al. (12) adopted the nearest neighbor regularization to eliminate the noise information in the metabolite-disease interaction network, and obtained good prediction results, which proved the effectiveness of the local structure information in the prediction of metabolite-disease interaction. Zhou et al. (14) achieved the accurate classification of unlabeled nodes by introducing local neighbor information. The construction strategy of the nearest neighbor graph determines the algorithm's performance. The nearest neighbor constraint usually adopts Laplacian graph regularization. However, Wang et al. (15) designed the local spectral matrix, called Vicus, which can outperform the Laplacian matrix in some scenarios.

In addition, LMF (logical matrix factorization) has been successfully applied in the biological interaction prediction. Johnson (16) demonstrated the advantages of logical matrix factorization in modeling unobserved connections, which was realized by setting different weights for positive and negative samples. Liu et al. (17) predicted the drug-target interaction by combining the neighbor structure of nodes and the logical matrix factorization algorithm.

In this paper, we propose a novel algorithm based on logical matrix factorization and considering the local structure information (using the aforementioned spectral matrix) to predict metabolite-disease interactions. The paper's main contributions are as follows.

(i) Integrating multisource information, such as disease description information from medical subject headings (MeSH) and disease-gene interaction information to build a disease similarity network. Multi-source information fusion can avoid the unreliability and inaccuracy in results caused by measurement errors and noises from a single data source, and it can describe the correlation between nodes more comprehensively;
(ii) The impact of noise and outliers is largely eliminated by employing the logical matrix factorization and local neighbor structure information. The neighbor's matrix constructed by the label diffusion algorithm has obvious advantages over the traditional Laplacian matrix. The experimental results show that the proposed method was superior to the baseline and state-of-the-art algorithms on the metabolite-disease dataset. The performance was improved by 5.28 and 5.61% in AUPR and F1, respectively;
(iii) The proposed method is easily extended to other biological problems, such as phage-host interaction prediction and metabolite-drug interaction prediction.

2. Materials and methods

2.1. Dataset

The collected data fall into three categories:

(i) Disease-related data, which were downloaded from the Comparative Toxicogenomic Database (CTD) (18). Data sources include: ① the human disease medical dictionary, which consists of 12,988 disease names, MeSH ID, Online Mendelian Inheritance in Man (OMIM) ID, disease synonyms, and the tree-structured disease representation; ② 25,114,553 interactions between 46,045 genes and 7,163 diseases; ③ 1,727,119 interactions between 13,126 Gene Ontology Biological Processes (GO BPs) and 7,116 diseases;

(ii) Metabolite-related data, which were collected from the Human Metabolome Database (HMDB) (19). The data include 814,427 interactions between 5,643 genes and 24,444 metabolites. Furthermore, the functional similarity network of metabolites was derived from the human gene interaction network (1);

(iii) Metabolite-disease interaction data, which were also obtained from the HMDB (19). Originally, the data contained 24,722 interactions between 649 diseases and 22,265 metabolites. By removing diseases without OMIM ID and semantic similarity and metabolites lacking functional similarity, we shrank that figure to 3,360 interactions between 337 diseases and 1,444 metabolites.

2.2. Problem formalization

In this article, the set of metabolites is denoted by $M = \{m_i\}_{i=1}^{n}$, and the set of diseases is denoted by $D = \{d_j\}_{j=1}^{m}$, where n and m are the number of metabolites and diseases, respectively. The known metabolite-disease interactions are represented as an $n \times m$ binary matrix ($Y \in R^{n \times m}$), where $y_{ij} = 1$ if a metabolite (m_i) has been observed to interact with a disease (d_j); otherwise, $y_{ij} = 0$.This study aimed to solve the problem of predicting the interaction probability of a disease-metabolite pair, and it subsequently ranked the candidate disease-metabolite pairs based on these probabilities in descending order. Thus, the top-ranked pairs can be viewed as latent interactions.

2.3. Metabolite-disease interaction prediction process based on logical matrix factorization and local neighborhood constraints

The prediction process, as demonstrated in Figure 1, can be divided into three subprocesses:

(i) The disease-disease similarity network is constructed by integrating the disease-related data (disease-gene interactions, disease-GO interactions, and the MeSH tree). Similarly, the metabolite-metabolite similarity network is built from the metabolite-related data (gene–gene associations, metabolite-gene interactions. Due to its highly sparse and noisy, the metabolite-disease interaction data is smoothed via WKNNP (20).

(ii) The local spectral matrices of diseases and metabolites are computed based on the disease-disease similarity network and metabolite-metabolite network, respectively.

(iii) The metabolite-disease interaction probabilities are computed by feeding the modified metabolite-disease interaction matrix, metabolite local spectral matrix, and disease local spectral matrix into the proposed logical matrix factorization model based on the local nearest neighbor constraint (LMFLNC).

Two crucial steps in the prediction process need further explanations.

(i) Disease-disease similarity network construction.

To obtain the comprehensive and accurate similarity between diseases, multiple data source of diseases including disease MeSH descriptors, disease–GO biological process interaction networks and disease–gene interaction networks are integrated. We employs the MultiSourcDSim model presented in (21) to calculate the semantic similarity of diseases. Specifically, for MeSH descriptors (22), we firstly construct a directed acyclic graph (DAG) to describe the relationships between any two diseases. Secondly, the probability of a disease term is calculated with its frequency occurring in the association dataset (Eqs. 1–2). Finally, the disease similarity (Eq. 3) is calculated with Lin's method (23).

$$f(t) = self(t) + \sum_{tc \in children(t)} f(tc) \tag{1}$$

$$prob(t) = \frac{f(t)}{N}. \tag{2}$$

$$score(t_1, t_2) = \max_{t \in LCA(t_1, t_2)} \left(\frac{2 \times \log prob(t)}{\log prob(t_1) + \log prob(t_2)} \right) \tag{3}$$

where $self(t)$ is the number of disease term t, tc is a direct child of t. $f(t)$ is the frequency at which t occurs in the single association dataset. N is the frequency of the root node term. $LCA(t_1, t_2)$ is the set of least common ancestors of term t_1 and t_2. $score(t_1, t_2)$ denotes the semantic similarity score between disease terms t_1 and t_2. For the other two data source, the similarity score is used to compute the disease similarity network.

(ii) Metabolite-metabolite similarity network construction.

With metabolite-gene interaction data, the similarity between any two genes, g_i and g_j, can be measured as

$$Sim\_g(g_i, g_j) = |GO_i \cap GO_j| / |GO_i \cup GO_j|, \tag{4}$$

where GO_i and GO_j denote the GO sets explaining g_i and g_j, respectively.

Similarly, the similarity between a gene (g_i) and a gene set (G) can be defined as

$$SG(g_i, G) = \max_{g_j \in G} \left[Sim\_g(g_i, g_j) \right]. \tag{5}$$

According to (24), the similarity between two metabolites, m_1 and m_2, can be computed as

$$SM(m_1, m_2) = \frac{\sum_{g_1 \in G_1} SG(g_1, G_2) + \sum_{g_2 \in G_2} SG(g_2, G_1)}{|G_1| + |G_2|}, \tag{6}$$

where G_1 and G_2 stand for gene sets related to m_1 and m_2, respectively; $|\bullet|$ denotes the set size.

The metabolite-metabolite similarity network is built via Equation (6).

2.4. Logical matrix factorization based on local nearest neighbor constraint

Logical matrix factorization has been successfully applied to the prediction of drug-target and virus-host interactions. In this

FIGURE 1
Flowchart of metabolite-disease interaction prediction with the LMFLNC. In step1, the disease-disease similarity matrix *A* is constructed by integrating the disease-related data disease-gene interactions, disease-GO interactions, and the MeSH tree with clusDCA. Similarly, the metabolite-metabolite similarity matrix *B* is built from gene–gene associations, metabolite-gene interactions; Running WKNNP on the metabolite-disease interaction matrix *C* generates the completed metabolite-disease interaction matrix *D*; In step 2, the local spectral matrices of diseases and metabolites *E* and *F* are obtained based on the disease-disease similarity matrix and metabolite-metabolite matrix, respectively. Finally, in step 3, the proposed LMFLNC was used to predict the metabolite-disease interaction scores.

paper, a new model based on logical matrix factorization is proposed to predict the interaction between metabolites and diseases. The main idea is to map metabolites and diseases into a shared low-dimensional latent semantic space, $r << \min(n,m)$. Then, the probability of interaction between metabolite m_i and disease d_j can be modeled by the following logical function:

$$\mathbf{p}_{ij} = \frac{\exp\left(\mathbf{u_i}\mathbf{v_j}^{\mathrm{T}}\right)}{1+\exp\left(\mathbf{u_i}\mathbf{v_j}^{\mathrm{T}}\right)}, \tag{7}$$

where $\mathbf{u_i} \in \mathrm{R}^{1\times r}$ and $\mathbf{v_j} \in \mathrm{R}^{1\times r}$ are latent representations of metabolite m_i and disease d_j, respectively.

In logical matrix factorization, the known or experimentally verified interactions are usually more informative, so they are usually assigned higher weights than those unknown ones. Each metabolite-disease interaction is regarded as $c(c \geq 1)$ positive sample, and each unknown metabolite-disease pair is regarded as a single negative sample. c is used to control the importance level of the observed interactions, which was empirically set to 2 in the subsequent experiments.

Assuming that each training sample is independent, according to the maximum likelihood estimation, the following probability representation can be obtained:

$$p(\mathbf{A}|\mathbf{U},\mathbf{V}) = \prod_{1\leq i\leq m,1\leq j\leq n,\mathbf{A}_{ij}=1}\left[\mathbf{p}_{ij}^{\mathbf{A}_{ij}}\left(1-\mathbf{p}_{ij}\right)^{1-\mathbf{A}_{ij}}\right]^{c} \times \prod_{1\leq i\leq m,1\leq j\leq n,\mathbf{A}_{ij}=0}\mathbf{p}_{ij}^{\mathbf{A}_{ij}}\left(1-\mathbf{p}_{ij}\right)^{1-\mathbf{A}_{ij}}. \tag{8}$$

where $\boldsymbol{A}$ represents the known metabolite-disease interaction matrix; $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the decomposed the metabolite and disease latent semantic matrices, respectively; m is the number of metabolites; n is the number of diseases. The logarithm of $p(\mathbf{A}|,\mathbf{U}|,\mathbf{V})$ can be inferred by combining Equation (7) with Equation (8):

$$\log\left[p(\mathbf{A}|\mathbf{U},\mathbf{V})\right]=\sum_{i=1}^{m}\sum_{j=1}^{n}c\mathbf{A}_{ij}\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}-\left(1+c\mathbf{A}_{ij}-\mathbf{A}_{ij}\right)\ln\left[1+\exp\left(\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}\right)\right]. \quad (9)$$

Equation (9) is also called the basic LMF objective function. The latent representation matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be estimated by maximizing this function.

To improve the performance of the logical matrix factorization algorithm, researchers (12, 17) introduced the local neighbor constraint. They sorted the nodes by their similarities to find neighbor nodes, but they ignored the diffusion and propagation of label information carried by neighbor nodes, which limited the performance enhancement. In this study, inspired by the idea of a local spectral matrix, the Vicus matrix (15), we obtained the following objective function by using the Vicus matrix to constrain Equation (9):

$$\log\left[p(\mathbf{A}|\mathbf{U},\mathbf{V})\right]=\sum_{i=1}^{m}\sum_{j=1}^{n}c\mathbf{A}_{ij}\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}-\left(1+c\mathbf{A}_{ij}-\mathbf{A}_{ij}\right)\ln\left[1+\exp\left(\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}\right)\right]+\lambda/2\left[tr\left(\mathbf{U}^{\mathrm{T}}\mathbf{vir}_u\mathbf{U}\right)+tr\left(\mathbf{V}^{\mathrm{T}}\mathbf{vir}_v\mathbf{V}\right)\right] \quad (10)$$

where λ is the regularization parameter to balance between the factorization error and the local spatial structure preservation; $\mathbf{vir}_u$ and $\mathbf{vir}_v$ represent the local spectral matrices of metabolites and diseases, respectively, whose calculation process is as follows:

Let $X=\{x_1,x_2,\cdots,x_n\}$ be the set of data points, $\mathbf{W}$ be the weighted network constructed from X with X as the vertex set and the similarities among X as the weight set; x_i be the ith data point in X; the ith vertex in $\mathbf{W}$, $N(i)$ be the neighbors of x_i, whose size is K; and C be the number of clusters.

First, for node x_i, subnet $\mathbf{W}_i(ver_i,\varepsilon_i)$ is extracted from $\mathbf{W}$, where the vertex set $ver_i=N(i)\cup x_i$, and ε_i is the edge set. Through the label diffusion algorithm (14), the label indicator vector is reconstructed as

$$\mathbf{F}_{ver_i}^{k}=(1-\alpha)(\mathbf{I}-\alpha\mathbf{S}_i)^{-1}\mathbf{q}_{ver_i}^{k},1\le k\le C, \quad (11)$$

where α is a constant between 0 and 1, which is set to 0.9, as suggested in (24); $\mathbf{q}_{ver_i}^{k}$ is the clustering indicator vector reflecting the scaling of subnet $\mathbf{W}_i$; $\mathbf{S}_i$ represents the standardized transition matrix of $\mathbf{W}_i$, defined as $\mathbf{S}_i(u,t)=\mathbf{W}_i(u,t)/\sum_{l=1}^{K+1}\mathbf{W}_i(u,l)$.

Second, $\mathbf{q}_{ver_i}^{k}$ is estimated by $\boldsymbol{F}_{ver_i}^{k}$. Let $\bar{\mathbf{q}}_{ver_i}^{k}=\mathbf{F}_{ver_i}^{k}[K+1]$ indicate the likelihood that data point i belongs to cluster k. The next task is to maximize the concordance between $\bar{\mathbf{q}}_{ver_i}^{k}$ and $\mathbf{q}_{ver_i}^{k}$. Let $\bar{\mathbf{q}}_{ver_i}^{k}=\beta_i\mathbf{q}_{ver_i}^{k}$, where $\beta_i\in\mathbf{R}^{K+1}$ is the row of matrix $(1-\alpha)(\mathbf{I}-\alpha\mathbf{S}_i)^{-1}$, which represents the convergence state of label diffusion. Thus, $\bar{\mathbf{q}}_{ver_i}^{k}$ can be estimated as

$$\bar{\mathbf{q}}_{ver_i}^{k}\approx\frac{\beta_i[1:K]\mathbf{q}_{N(i)}^{k}}{1-\beta_i[K+1]}, \quad (12)$$

where $\beta_i[1:K]$ and $\beta_i[K+1]$ denotes the first K elements and the $(K+1)$th element in β_i, respectively.

Afterward, matrix $\boldsymbol{B}$ is constructed to represent the linear relationship between $\mathbf{q}^k$ and $\bar{\mathbf{q}}_k{:}\bar{\mathbf{q}}_k=\mathbf{B}\mathbf{q}^k$. It is computed as

$$\mathbf{B}_{ij}=\begin{cases}\dfrac{\beta_{ij}}{1-\beta_i[K+1]}, & x_j\in N(i)\\ 0, & \text{otherwise}\end{cases}. \quad (13)$$

To minimize the difference between $\mathbf{q}^k$ and $\bar{\mathbf{q}}_k$, we can define an objective function as

$$\sum_{i=1}^{n}\sum_{k=1}^{C}\left(\bar{\mathbf{q}}_i^k-\mathbf{q}_i^k\right)^2=\sum_{k=1}^{C}\mathbf{q}^k-\bar{\mathbf{q}}^{k2}\approx\sum_{k=1}^{C}\mathbf{q}^k-\mathbf{B}\mathbf{q}^{k2}=Trace\left[\mathbf{Q}^{\mathrm{T}}(\mathbf{I}-\mathbf{B})^{\mathrm{T}}(\mathbf{I}-\mathbf{B})\mathbf{Q}\right] \quad (14)$$

Finally, let $\mathbf{vir}=(\mathbf{I}-\mathbf{B})^{\mathrm{T}}(\mathbf{I}-\mathbf{B})$, which is the needed local spectral matrix. Wang et al. (15) proved that $\mathbf{vir}$ and the Laplacian matrix share many of the same properties. For example, they are both symmetric and positive semidefinite, with the minimum eigenvalue being 0 and the eigenvector being 1.

In logical matrix factorization, to prevent overfitting, we usually constrain the latent space matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ to construct the final objective function as

$$\log\left[p(\mathbf{A}|\mathbf{U},\mathbf{V})\right]=\sum_{i=1}^{m}\sum_{j=1}^{n}c\mathbf{A}_{ij}\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}-\left(1+c\mathbf{A}_{ij}-\mathbf{A}_{ij}\right)\ln\left[1+\exp\left(\mathbf{u}_i\mathbf{v}_j^{\mathrm{T}}\right)\right]+\alpha/2\left(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2\right)+\lambda/2\left[tr\left(\mathbf{U}^{\mathrm{T}}\mathbf{vir}_u\mathbf{U}\right)+tr\left(\mathbf{V}^{\mathrm{T}}\mathbf{vir}_v\mathbf{V}\right)\right] \quad (15)$$

where α denotes the regularization parameter. Here, the gradient descent algorithm is used to optimize Equation (15). Specifically, let L represent the objective function whose partial derivatives with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ are given as follows:

$$\frac{\partial L}{\partial\mathbf{U}}=\mathbf{P}\mathbf{V}+(c-1)(\mathbf{A}\odot\mathbf{P})\mathbf{V}-c\mathbf{A}\mathbf{V}+(\alpha I+\lambda\mathbf{vir}_{\mathbf{u}})\mathbf{U}, \quad (16)$$

$$\frac{\partial L}{\partial\mathbf{V}}=\mathbf{P}^{\mathrm{T}}\mathbf{U}+(c-1)\left(\mathbf{A}^{\mathrm{T}}\odot\mathbf{P}^{\mathrm{T}}\right)\mathbf{U}-c\mathbf{A}^{\mathrm{T}}\mathbf{U}+(\alpha I+\lambda\mathbf{vir}_{\mathbf{v}})\mathbf{V}, \quad (17)$$

where $\boldsymbol{P}$ is the probability matrix defined by Equation (7), and $\odot$ represents the Hadamard product of a matrix. After the latent representations of $\boldsymbol{U}$ and $\boldsymbol{V}$ have been acquired, any unknown metabolite-disease interaction probability can be predicted by Equation (7). However, in the training process, the latent vectors of some unobserved metabolites and diseases are obtained based on negative samples, which may not

be accurate enough. Ma et al. (12) presented an effective solution. Let $N_m^+ = \left\{ m_i | \sum_j \mathbf{A}_{ij} > 0 \right\}$ represent the set of metabolites interacting with any disease, and let $N_m^+(m_i)$ represent the set of K nearest neighbors of metabolites in N_m^+. We set $K = 10$ in this manuscript. Metabolite m_i can be represented by a linear combination of the latent vectors of $N_m^+(m_i)$, and is defined as follows:

$$\bar{\mathbf{u}}_i = \begin{cases} \mathbf{u}_i, & m_i \in N_m^+ \\ 1/Q_{m_i} \sum_{k=1}^{K} w_k^m \mathbf{u}_k, & m_i \notin N_m^+ \end{cases}, \quad (18)$$

where $Q_{m_i} = \sum_{k=1}^{K} \alpha^{k-1} \overline{\mathbf{ConsM}}(m_i, N_i^k)$ is a normalized term, $N_i^k \in N_m^+(m_i)$ denotes the kth neighbor of m_i, and $\overline{\mathbf{ConsM}}$ is the binary neighbor similarity matrix. $\overline{\mathbf{ConsM}}_{ij} = \mathbf{ConsM}_{ij}$ if metabolite$m_i \in N(m_j)$ or metabolite $m_j \in N(m_i)$; otherwise, $\overline{\mathbf{ConsM}}_{ij} = 0$. $\alpha \in [0,1]$ is a decay factor, which is 0.9 in this paper. $w_k^m = a^{k-1} \overline{\mathbf{ConsM}}(m_i, N_i^k)$. Similarly, the representation of disease d_j can be obtained:

$$\bar{\mathbf{v}}_j = \begin{cases} \mathbf{v}_j, & d_j \in N_d^+ \\ 1/Q_{d_j} \sum_{k=1}^{K} w_k^d \mathbf{v}_k, & d_j \notin N_d^+ \end{cases}, \quad (19)$$

where $Q_{d_j} = \sum_{k=1}^{K} \alpha^{k-1} \overline{\mathbf{ConsD}}(d_j, N_j^k)$ is a normalized term, N_j^k denotes the kth neighbor of disease d_j, and $w_k^d = a^{k-1} \overline{\mathbf{ConsD}}(d_j, N_j^k)$

Eventually, the probability of an interaction between metabolite m_i and disease d_j can be rewritten as

$$p(m_i, d_j) = \frac{\exp(\bar{\mathbf{u}}_i \bar{\mathbf{v}}_j^{\mathrm{T}})}{1 + \exp(\bar{\mathbf{u}}_i \bar{\mathbf{v}}_j^{\mathrm{T}})}. \quad (20)$$

In order to clearly demonstrate the steps of LMFLNC algorithm, we also presented its pseudocode in Table 1.

3. Results and discussion

3.1. Experimental settings and evaluation metrics

Following the previous studies, we used the fivefold cross-validation technique for model validation in this paper. In each round, one-fifth of the known metabolite-disease interactions and all unobserved interactions (metabolite-disease pairs corresponding to elements of value 0 in the metabolite-disease interaction matrix A) were used for testing; the rest were used for training. AUPR, AUC, and F1 were adopted as performance evaluation. To achieve a relatively objective evaluation, we randomly ran the cross validation 20 times, over which the average values of the aforementioned metrics were taken as their final values. The model implementation and validation were realized in MATLAB R2017b (see Table 1).

3.2. Experimental results

To verify the superiority of the proposed LMFLNC model, we compared it with such baselines as MN-LMF (12), PROFANCY (2), WMAN (25), and MCF (13). The parameters of PROFANCY, WMAN, and MN-LMF were set to default values. For MCF, the reboot probability is set as the optimal element from $\{0.1, 0.2, \cdots, 0.9\}$. For LMFLNC, we set the number of nearest neighbors in local spectral matrices of metabolites and diseases as $K = 15$, the importance level of observed interactions $c = 2$, the neighbor regularization parameter $\lambda = 8$, and the latent space regularization parameter $\alpha = 4$. The performance of the abovementioned algorithms on the metabolite-disease benchmark dataset is shown in Table 2.

Table 2 shows that the LMFLNC algorithm outperformed the second MN-LMF algorithm in AUPR and F1 5.28 and 5.61%, respectively. Additionally, the prediction performances of WMAN and MCF methods were unsatisfactory. One possible reason is that these two methods simply focus on the known metabolite-disease interaction network and only leverage limited prior knowledge, that is, the disease similarity network. However, the LMFLNC method fully considers the similarities of metabolites and diseases at multiple levels and then adjusts the importance level of positive and negative samples (the observed metabolite-disease interaction is regarded as a positive sample. The unobserved metabolite-disease interaction is regarded as a negative sample) by parameter c, which improved its performance. Moreover, compared with MN-LMF, LMFLNC uses the local spectral matrices of metabolites and diseases to construct neighbor constraints, so the latent representations of metabolites and diseases generated by the logical matrix factorization were more robust. The experimental results show the potential of LMFLNC in predicting unknown metabolite-disease interactions.

3.3. Parameter analysis

Two parameters need to be tuned in LMFLNC: the latent space regularization parameter α and the local spectral parameter (or neighbor regularization parameter) λ; the other ones are set by default. The grid search was employed to find the optimal parameter values. Let $\alpha \in \{2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3\}$, $\lambda \in \{2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3\}$, and the model performance over different parameter combinations

TABLE 1 The pseudocode of the LMFLNC algorithm.

Input: The metabolite-disease interaction matrix A; parameters c, α, λ
Output: The latent representation matrices, U and V
1. Calculate the disease–disease similarity matrix using Eq. (3); Calculate the metabolite–metabolite similarity matrix according to Eq. (6);
2. Calculate the spectral matrices of metabolites and diseases;
3. Calculate the modified metabolite-disease interaction matrix via WKNNP (20);
4. Initialize U and V randomly;
5. For $t = 1, \ldots\ldots, max\_iter$ do
6. Update U and V according to Adara algorithm
7. Until convergence conditions are satisfied
8. End for
9. Return U, V

was evaluated by a fivefold cross-validation. As shown in Figure 2, LMFLNC obtained the optimal prediction performance (AUPR) when $\alpha = 4$ and $\lambda = 8$.

3.4. Case studies

We further verify the performance of LMFLNC method in this section. First, the entire dataset was used to train LMFLNC with the optimal parameters obtained above. Then, the trained LMFLNC was used to predict the interaction probabilities between all the metabolites and two example diseases, "21-Hydroxylase deficiency" and "3-Hydroxy-3-methylglutaryl-CoA lyase deficiency," in the dataset. Table 3 displays 10 metabolites relating to the first example disease, with the probabilities listed in descending order. Similarly, Table 4 displays 15 metabolites relating to the second example disease, with the probabilities again listed in descending order.

It can be seen that all of the nine metabolites related to the disease "21-Hydroxylase deficiency" in the dataset appear in Table 3 and, more importantly, are located in the top nine. Similarly, all of the 13 metabolites related to disease "3-Hydroxy-3-methylglutaryl-CoA lyase deficiency" in the dataset are included in Table 4 and occupy the top 13. These findings demonstrate the good accuracy of LMFLNC. Note that LMFLNC also predicted that the metabolite 'Cortisol "(HMDB0000063) were likely to interact with disease" "21-Hydroxylase deficiency" (the likelihood is 0.5896) and that metabolites "3-Hydroxybutyric acid (HMDB0000011)" and "Acetoacetic acid (HMDB0000060)" were likely to interact with disease "3-Hydroxy-3-methylglutaryl-CoA lyase deficiency" (likelihood of 0.4991 and 0.3614, respectively). Two of these three predictions have been verified, showing the potential of the LMFLNC model to discover latent metabolite-disease interactions.

In the same way, LMFLNC can compute the probabilities of diseases relating to a specific metabolite and predict new disease-metabolite interactions.

TABLE 2 Performance comparison of metabolite-disease benchmark dataset.

Algorithm	AUPR	AUC	F1
WMAN	0.0151	0.6181	0.0800
PROFANCY	0.2325	0.9027	0.3066
MCF	0.0151	0.6156	0.0770
MN-LMF	0.3731	0.9659	0.4135
LMFLNC	0.3931	0.9661	0.4367

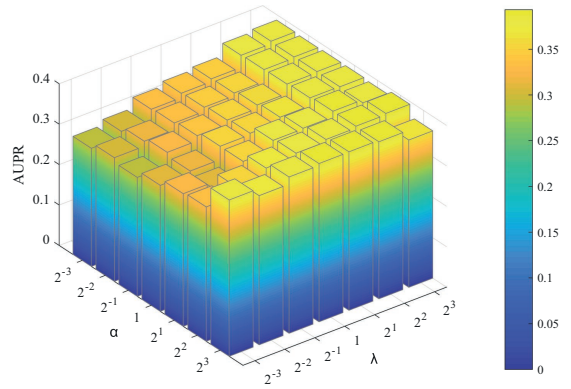

FIGURE 2
Parameter sensitivity analysis. AUPR achieves the maximum value when α =4 and λ=8.

4. Conclusion

Existing metabolite-disease interaction prediction methods mainly leverage the global similarity network, which may be limited by noise and outliers. To solve this problem, we introduced a novel method, LMFLNC, to predict the metabolite-disease interaction. Extensive experiments were conducted on the collected dataset. The results show that the proposed LMFLNC method outperformed the baselines. LMFLNC also revealed several potential metabolite-disease interactions, such as "Cortisol (HMDB0000063)," relating to "21-Hydroxylase deficiency," and "3-Hydroxybutyric acid (HMDB0000011)" and "Acetoacetic acid (HMDB0000060)," both relating to "3-Hydroxy-3-methylglutaryl-CoA lyase deficiency."

Despite its promising performance, LMFLNC has the following weaknesses. (1) The predicted new metabolite-disease interactions need further verification. (2) The dataset scale,

TABLE 3 21-Hydroxylase deficiency' related metabolites (top 10, descend).

NO.	Metabolite ID	Metabolite name	Interaction probability	Category
1	HMDB0000374	17-Hydroxyprogesterone	0.9568	Known
2	HMDB0000053	Androstenedione	0.9308	Known
3	HMDB0000122	D-Glucose	0.9279	Known
4	HMDB0000234	Testosterone	0.9266	Known
5	HMDB0000586	Potassium	0.9241	Known
6	HMDB0000595	Hydrogen carbonate	0.9224	Known
7	HMDB0000588	Sodium	0.9056	Known
8	HMDB0000077	Dehydroepiandrosterone	0.8155	Known
9	HMDB0004030	21-Deoxycortisol	0.7693	Known
10	HMDB0000063	Cortisol	0.5896	PubMed:16439592

TABLE 4 3-Hydroxy-3-methylglutaryl-CoA lyase deficiency-related metabolites (top 15, descend).

NO.	Metabolite ID	Metabolite name	Interaction probability	Category
1	HMDB0000122	D-Glucose	0.9776	Known
2	HMDB0000190	L-Lactic acid	0.9735	Known
3	HMDB0000051	Ammonia	0.9456	Known
4	HMDB0000754	3-Hydroxyisovaleric acid	0.9455	Known
5	HMDB0000661	Glutaric acid	0.9295	Known
6	HMDB0000062	L-Carnitine	0.9209	Known
7	HMDB0000243	Pyruvic acid	0.9152	Known
8	HMDB0000595	Hydrogen carbonate	0.8861	Known
9	HMDB0000063	Cortisol	0.8805	Known
10	HMDB0000357	3-Hydroxybutyric acid	0.8802	Known
11	HMDB0000459	3-Methylcrotonylglycine	0.8754	Known
12	HMDB0000355	3-Hydroxymethylglutaric acid	0.7789	Known
13	HMDB0000509	Senecioic acid	0.7646	Known
14	HMDB0000011	3-Hydroxybutyric acid	0.4991	PubMed:12072887
15	HMDB0000060	Acetoacetic acid	0.3614	Unconfirmed

including the data quantity and type, is relatively small, and the information of metabolite structure and pathway can be incorporated to improve the performance and robustness of LMFLNC.

Our future research work will include the following: (1) exploring combining multi-kernel learning and logical matrix factorization in a study on the metabolite-disease interaction relationship and (2) exploring the application of our model in similar fields, such as microorganism-drug interactions and microorganism-metabolite interactions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YZ and YM wrote the manuscript and developed the algorithms. YM designed the concept and including the structure and content of the manuscript. YM, QZ, and YZ critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by Hubei Superior and Distinctive Discipline Group of "New Energy Vehicle and Smart Transportation."

Acknowledgments

The authors thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Yao Q, Xu Y, Yang H, Shang D, Zhang C, Zhang Y, et al. Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Sci Rep*. (2015) 5:17201. doi: 10.1038/srep17201

2. Shang D, Li C, Yao Q, Yang H, Xu Y, Han J, et al. Prioritizing candidate disease metabolites based on global functional relationships between metabolites in the context of metabolic pathways. *PLoS One*. (2014) 9:e104934. doi: 10.1371/journal.pone.0104934

3. Nicholson JK, Lindon JC. Systems biology: metabonomics. *Nature*. (2008) 455:1054–6. doi: 10.1038/4551054a

4. Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. *Cells*. (2008) 134:714–7. doi: 10.1016/j.cell.2008.08.026

5. Ouyang D, Xu J, Huang H, Chen Z. Metabolomic profiling of serum from human pancreatic cancer patients using 1H NMR spectroscopy and principal component analysis. *Appl Biochem Biotechnol*. (2011) 165:148–4. doi: 10.1007/s12010-011-9240-0

6. Reinke SN, Gallart-Ayala H, Gomez C, Checa A, Fauland A, Naz S, et al. Metabolomics analysis identifies different metabotypes of asthma severity. *Eur Respir J*. (2017) 49:1601740. doi: 10.1183/13993003.01740-2016

7. Ibanez C, Simo C, Martin-Alvarez PJ, Kivipelto M, Winblad B, Cedazo-Minguez A, et al. Toward a predictive model of Alzheimer's disease progression using capillary electrophoresis-mass spectrometry metabolomics. *Anal Chem*. (2012) 84:8532–40. doi: 10.1021/ac301243k

8. Wang B, Chen D, Chen Y, Hu Z, Cao M, Xie Q, et al. Metabonomic profiles discriminate hepatocellular carcinoma from liver cirrhosis by ultraperformance liquid chromatography-mass spectrometry. *J Proteome Res*. (2012) 11:1217–27. doi: 10.1021/pr2009252

9. Lei X, Tie J. Prediction of disease-related metabolites using bi-random walks. *PLoS One*. (2019) 14:e0225380. doi: 10.1371/journal.pone.0225380

10. Wang Y, Juan L, Peng J, Zang T, Wang Y. Prioritizing candidate diseases-related metabolites based on literature and functional similarity. *BMC Bioinform*. (2019) 20:574. doi: 10.1186/s12859-019-3127-4

11. Mi K, Jiang Y, Chen J, Lv D, Qian Z, Sun H, et al. Construction and analysis of human diseases and metabolites network. *Front Bioeng Biotechnol*. (2020) 8:398. doi: 10.3389/fbioe.2020.00398

12. Ma Y, He T, Jiang X. Multi-network logistic matrix factorization for metabolite-disease interaction prediction. *FEBS Lett*. (2020) 594:1675–84. doi: 10.1002/1873-3468.13782

13. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci USA*. (2018) 115:7723–8. doi: 10.1073/pnas.1805681115

14. Zhou D, Bousquet O, Lal T, Weston J, Schölkopf B. Learning with local and global consistency. *Adv Neural Inf Proces Syst*. (2004) 16:16.

15. Wang B, Huang L, Zhu Y, Kundaje A, Batzoglou S, Goldenberg A. Vicus: exploiting local structures to improve network-based analysis of biological data. *PLoS Comput Biol*. (2017) 13:e1005621. doi: 10.1371/journal.pcbi.1005621

16. Johnson CC, Logistic matrix factorization for implicit feedback data, in NIPS workshop on distributed machine learning and matrix computations (2014).

17. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol*. (2016) 12:e1004760. doi: 10.1371/journal.pcbi.1004760

18. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Mcmorran R, Wiegers J, et al. The comparative Toxicogenomics database: update 2019. *Nucleic Acids Res*. (2019) 47:D948–54. doi: 10.1093/nar/gky868

19. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*. (2018) 46:D608–17. doi: 10.1093/nar/gkx1089

20. Xiao Q, Luo J, Liang C, Cai J, Ding P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics*. (2018) 34:239–8. doi: 10.1093/bioinformatics/btx545

21. Deng L, Ye D, Zhao J, Zhang J. MultiSourcDSim: an integrated approach for exploring disease similarity. *BMC Med Inform Decis Mak*. (2019) 19:269. doi: 10.1186/s12911-019-0968-8

22. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007. (1995).

23. Lin D. An information-theoretic definition of similarity. *Icml*. (1998): 296–304.

24. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. *Bioinformatics*. (2007) 23:1274–81. doi: 10.1093/bioinformatics/btm087

25. Hu Y, Zhao T, Zhang N, Zang T, Zhang J, Cheng L. Identifying diseases-related metabolites using random walk. *BMC Bioinform*. (2018) 19:116. doi: 10.1186/s12859-018-2098-1