Dual-Network Cross-Learning for Metabolite-Disease Association Prediction

Yanxin Chen, Qiao Ning, Yitong Zhang, Hui Li, and Shikai Guo

***Abstract*—In recent years, increasing evidence has demonstrated a close association between metabolites and various complex human diseases, providing valuable insights for disease diagnosis, treatment, and prevention. Although deep learning-based approaches have achieved certain success in predicting metabolic disease associations, challenges remain in enriching graph information and effectively integrating metabolic and disease features. To address these issues, this paper proposes a model named DCMDA, which extracts deep features of both metabolites and diseases using Dual-network Cross-learning for Metabolite-Disease Association prediction. DCMDA consists of three parts. The data processing module integrates similarity networks with association networks to construct a heterogeneous network. The feature extraction module extracts features from the metabolite-disease association network based on the non-negative matrix factorization method and from the heterogeneous network using graph autoencoder techniques. The feature fusion module combines the association matrix feature with the heterogeneous network feature through a Cross-Attention mechanism, thereby obtaining deep representations of metabolites and diseases. These features are then used to train the model to predict association scores between metabolites and diseases. Experimental results demonstrate that in 5-fold cross-validation, DCMDA achieves an area under the receiver operating characteristic curve (AUC) of 97.8% and an area under the precision-recall curve (AUPR) of 97.9%, outperforming state-of-the-art prediction methods.**

***Index Terms*—Metabolite-disease association, graph autoencoder, non-negative matrix factorization, Cross-Attention mechanism.**

I. Introduction

METABOLISM is a series of ordered chemical reactions that have significant impacts on the growth, reproduction, and response to external environments of organisms [1]. Substances produced or consumed during metabolic processes are referred to as metabolites. The levels of metabolites can directly reflect the physiological state of an organism, and ample evidence suggests that diseases are invariably accompanied by changes in metabolites. Therefore, in recent studies, numerous metabolites have been identified as features of diseases [2]. For instance, through extensive experiments and clinical cases, many metabolites with significant changes, such as blood glucose concentration, have gradually been recognized by physicians as one of the diagnostic criteria for diabetes in recent years [3]. There is an adverse association between circulating trimethylamine-N-oxide (TMAO) levels and the presence and severity of non-alcoholic fatty liver disease (NAFLD), while there is a favorable relationship between betaine and non-alcoholic fatty liver disease (NAFLD) [4]. Enrichment of Fusobacterium can induce metastasis of colorectal cancer [5]. Microbial metabolites in the gut are considered as new risk factors for cardiovascular disease and premature death [6]. The dysregulation of fatty acid metabolites plays a crucial role in the progression of complex diseases such as cardiovascular diseases, digestive system disorders, and metabolic diseases [7]. Choline is essential for normal liver, muscle, brain, lipid metabolism, cellular membrane composition, and repair [8].

Therefore, discovering the relationship between metabolites and diseases is crucial for medical researchers to gain a deeper understanding of the complex pathological processes of diseases. In recent years, many biologists have made significant efforts in researching metabolites and diseases. For example, Hori et al. [9] conducted metabolomic analyses of lung cancer patients using gas chromatography-mass spectrometry (GC-MS). Czech et al. [10] employed gas chromatography-liquid chromatography-tandem mass spectrometry (GC-MS and LC-MS/MS), combined with univariate and multivariate statistical analysis, to identify metabolic changes caused by the Alzheimer's disease phenotype. Bhattacharya et al. [11] performed mass spectrometry-based analyses, validating the independent associations of metabolites involved in branched-chain amino acid metabolism with extreme cases of coronary artery disease (CAD). However, due to limitations in time, funding, and accuracy, the efficiency of traditional biological experiments in achieving their goals is not satisfactory. Therefore, developing computational methods to efficiently and reliably explore

Received 19 June 2024; revised 19 December 2024; accepted 29 December 2024. Date of publication 16 January 2025; date of current version 3 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62302075, in part by Innovation Support Program for Dalian High Level Talents under Grant 2023RQ007, and in part by the Dalian Excellent Young Project under Grant 2022RY35. *(Corresponding authors: Qiao Ning; Shikai Guo.)*

Yanxin Chen, Yitong Zhang, and Hui Li are with the Department of Information Science and Technology, Dalian Maritime University, Dalian 116026, China.

Qiao Ning is with the Department of Information Science and Technology, Dalian Maritime University, Dalian 116026, China, also with School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China, and also with Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China (e-mail: ningq669@dlmu.edu.cn).

Shikai Guo is with the Department of Information Science and Technology, Dalian Maritime University, Dalian 116026, China, and also with the Dalian Key Laboratory of Artificial Intelligence, Dalian 116026, China (e-mail: Shikai.guo@dlmu.edu.cn).

The source code and data are available at https://github.com/ChenYanxin99/DCMDA.git.

Digital Object Identifier 10.1109/TCBBIO.2025.3527457

potential metabolite-disease associations is of great importance for human health and medical advancement.

In recent years, various computational methods have been introduced to predict associations between metabolites and diseases, primarily utilizing network algorithms and machine learning techniques. For instance, Yang et al. [12] used random walks to identify disease-associated metabolites, marking the initial application of computational methods in predicting metabolite-disease associations. However, the dataset used in their study contained a limited number of metabolite types, and the similarity matrices were sparse, as they only considered metabolite similarity based on disease similarity. Subsequently, Cheng et al. applied the KATZ [13] algorithm to predict associations between metabolites and diseases, marking the first use of the KATZ algorithm in metabolomics. Network algorithms often rely on small datasets, which may not effectively handle sparse data and noise present in the model.

To improve prediction accuracy, machine learning algorithms have been employed for metabolite-disease association prediction. For example, Lei et al. [14] used relationship completion-based non-negative matrix factorization to predict associations between metabolites and diseases. Zhang et al. developed the LightGBM [15] algorithm, which is based on gradient boosting, for metabolite-disease association prediction.

Benefiting from advancements in machine learning, many recent studies have utilized deep learning methods for metabolite-disease association prediction. Hierarchical structures based on deep learning models can facilitate abstract learning and capture subtle features of metabolites and diseases. In 2022, Sun et al. proposed a deep learning method based on graph neural networks (GCNAT) [16] for identifying disease-related metabolites. In the same year, Tie et al. introduced a metabolite-disease association prediction algorithm based on DeepWalk and random forests [17]. In 2023, Gao et al. [18] developed a deep learning model combining autoencoders and non-negative matrix factorization to extract features of metabolites and diseases, followed by training with multilayer perceptrons. Among all computational methods currently available, deep learning algorithms are the most effective. However, previous deep learning-based methods still face challenges, such as insufficiently rich graph information and the inability to effectively fuse metabolite and disease features. Developing computational models with high accuracy and robust performance for predicting potential metabolite-disease relationships remains a challenge.

Accordingly, this paper proposes a novel deep learning-based method named DCMDA for metabolite-disease association prediction using dual-network cross-learning. DCMDA extracts association matrix features using Non-negative Matrix Factorization (NMF) from the metabolite-disease association network. To enrich graph information, DCMDA integrates three types of association information: metabolite-metabolite, disease-disease, and metabolite-disease, to construct a heterogeneous graph, which is utilized to extract heterogeneous network features using a Graph Autoencoder with an Attention mechanism (GATE). Heterogeneous network features and association matrix features have their own specific characteristics. Therefore, DCMDA employs a Cross-Attention mechanism to enable the heterogeneous network features to focus only on key information within the association matrix features, rather than all information. This approach fuses metabolite and disease features, resulting in a deep representation of diseases and metabolites. The main contributions of this paper are summarized as follows:

1) We designed a data preprocessing module that integrates different similarity networks of metabolites and diseases using nonlinear methods, and combined them with the metabolite-disease association network to form a heterogeneous graph. This approach captures shared and complementary information from various data sources more effectively, addressing the issue of insufficiently rich graph information.
2) In addition to extracting the correlation matrix features through the NMF module, we further use the GATE module to extract complex network features from heterogeneous maps to enrich the feature information.
3) Through Cross-Attention, we cross-fuse the association matrix features with the heterogeneous network features, focusing on key features to obtain deep representations of diseases and metabolites. This effectively eliminates the influence of noise and resolves the issue of inadequate integration of metabolite and disease features in current deep learning methods.

II. DCMDA Model

The workflow diagram of DCMDA is illustrated in Fig. 1. DCMDA is primarily divided into three modules: the data processing module, the feature extraction module (comprising NMF and GATE), and the feature fusion module. In the data processing stage, the semantic similarity matrices, Gaussian kernel similarity matrices, and information entropy similarity matrices of diseases, as well as the structural, Gaussian kernel, and information entropy similarity matrices of metabolites, are normalized. Using the same nonlinear method, the three similarity matrices for diseases and the three similarity matrices for metabolites are integrated into two fused similarity networks, which are then combined with the metabolite-disease association matrix to form a heterogeneous graph. In the feature extraction module, NMF is used to extract association matrix features from the metabolite-disease association network, and GATE is applied to extract heterogeneous network features from the heterogeneous graph. Subsequently, the two sets of features for metabolites and diseases, along with their corresponding feature vectors and labels, are sent to the feature fusion module. Through a Cross-Attention mechanism, these features are fused and used for training and prediction.

A. Data Preprocessing Module

1) Datasets: The dataset used in this study was sourced from the Human Metabolome Database (HMDB, https://hmdb.ca/) [18], an online resource providing detailed information on metabolites within the human body. We selected 4,536 metabolite-disease pairs from HMDB, involving 2,262 metabolites and 216 diseases.

Unknown and unrelated metabolite-disease pairs are referred to as negative samples, while known metabolite-disease associations are referred to as positive samples. Due to the substantially

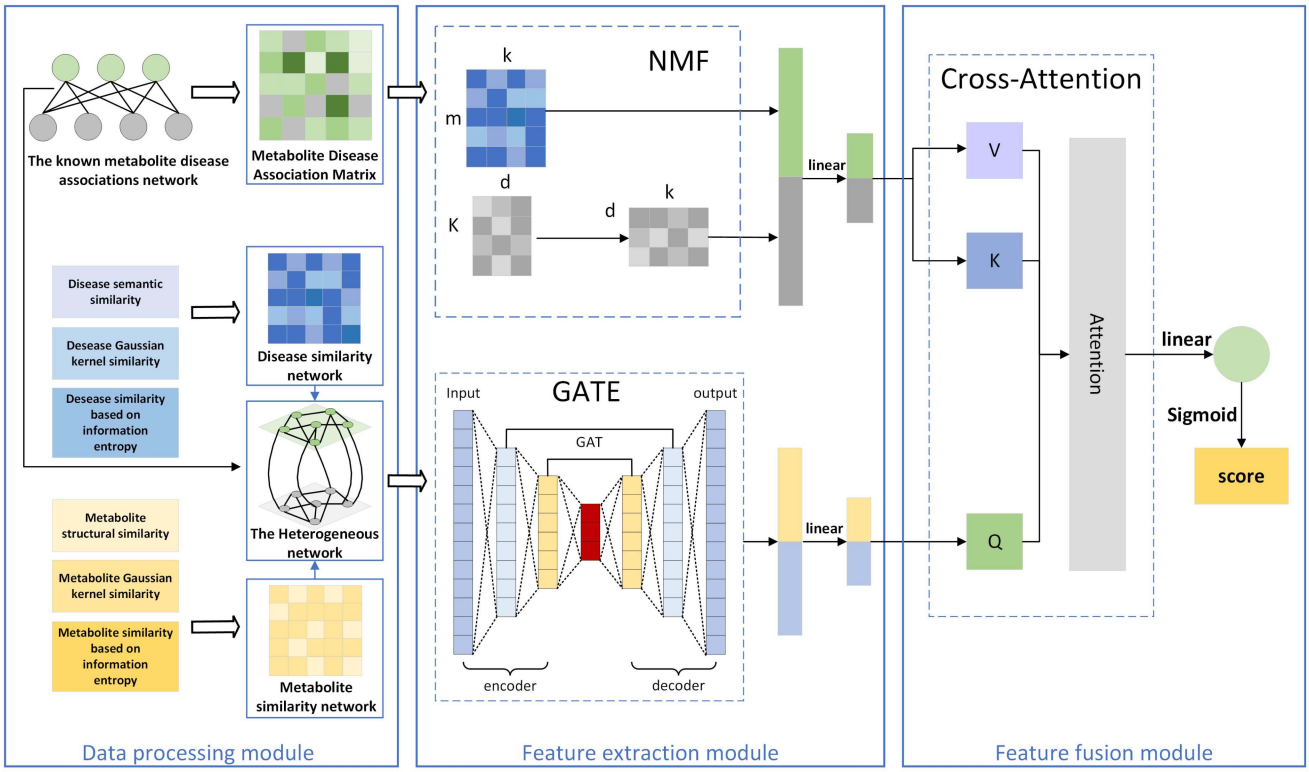

Fig. 1. The framework of DCMDA.

larger number of negative samples compared to positive samples, we randomly selected an equal number of negative samples (4,536) to match the number of positive samples. The positive samples and the randomly selected negative samples were then combined to form a new metabolite-disease association dataset for subsequent experiments.

2) Network Fusion Process of Similarity: Similar diseases typically exhibit comparable biological characteristics and pathogenic mechanisms. Therefore, similar diseases are likely to be associated with the same metabolites. Likewise, similar metabolites may also be linked to the same diseases. Metabolite-disease association data can be affected by experimental techniques and environmental factors, leading to the presence of noise. Introducing similarity information can better capture shared and complementary insights from multiple data sources to reduce the impact of noise, thereby improving the quality and reliability of the data. Similarity information can also serve as additional features, providing more context to comprehensively reveal potential associations between metabolites and diseases. By integrating various similarity networks into a fused similarity network, we can account for similarity features at different levels, evaluate the similarity relationships between metabolites and diseases from multiple perspectives, reduce the model's dependence on single sources of similarity information, and thus enhance the robustness of the model.

Our similarity fusion networks consist of three similarity matrices each. We calculated various similarity networks following the methods described in the references. The disease similarity fusion network comprises the semantic similarity of diseases [49], Gaussian kernel similarity of diseases [16], and disease information entropy similarity [50]. The metabolite similarity fusion network comprises the structural similarity of metabolites [18], Gaussian kernel similarity of metabolites [16], and metabolite information entropy similarity [50]. We employ a nonlinear method to integrate different similarity networks for metabolites and diseases. Similar to the integration method for metabolite similarity networks, we elaborate on the integration method using diseases as an example:

To ensure similar reference standards when integrating different networks, we subject all networks before integration to the same normalization process, ensuring that each node in the network has similar weights after normalization. Taking the semantic similarity network of diseases as an example, the normalization process is as follows, as depicted in the following formula:

$$SD_{DSS}(i,j) = \begin{cases} \frac{DSS(i,j)}{2*\sum_{k\neq i} DSS(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases} \quad (1)$$

The disease semantic similarity network, denoted as DSS, undergoes normalization by setting all diagonal elements to 1/2, and ensuring that the sum of each row's elements equals 1. This process yields the normalized disease semantic similarity network SD_{DSS}. Similarly, we can obtain the normalized disease Gaussian kernel similarity network SD_{DGIP} and the normalized disease information entropy similarity network SD_{DSIE} using the same method.

Next, we utilize K-nearest neighbors (KNNs) to compute the local affinity of disease semantic similarity between disease i

and j, denoted as local affinity S_kn_{DSS}.

$$S\_kn_{DSS}(i,j) = \begin{cases} \frac{DSS(i,j)}{\sum_{k \in Ni} DSS(i,k)}, & j \in Ni \\ 0, & otherwise \end{cases} \quad (2)$$

Following the principle that proximity implies higher similarity, K-nearest neighbors (KNN) first identifies its K nearest neighbor nodes, which are the K nodes closest to the given node in feature space. Here, N_i equals the total number of metabolites divided by 10, defining the KNN set for the given node. For remote nodes distant from the given node, their similarity scores are set to 0. Similarly, we can obtain the local affinity S_kn_{DGIP} of GIP between diseases i and j, as well as the local affinity S_kn_{DSIE} of SIE between diseases i and j, using the aforementioned approach.

Subsequently, utilizing the calculated normalized network and local affinity, each similarity network is iteratively updated.

$$SD_{DSS}^{(t)} = S\_kn_{DSS} \times (S\_kn_{DSS})^T \times (SD_{DGIP}^{(t-1)} + SD_{DSIE}^{(t-1)})/2 \quad (3)$$

$$SD_{DGIP}^{(t)} = S\_kn_{DGIP} \times (S\_kn_{DGIP})^T \times (SD_{DSS}^{(t-1)} + SD_{DSIE}^{(t-1)})/2 \quad (4)$$

$$SD_{DSIE}^{(t)} = S\_kn_{DSIE} \times (S\_kn_{DSIE})^T \times (SD_{DSS}^{(t-1)} + SD_{DGIP}^{(t-1)})/2 \quad (5)$$

In each iteration, the change in SD between consecutive steps is computed. If this change is sufficiently small, the algorithm is considered to have converged to a stable state, and the iteration ends. The convergence criterion formula is expressed as follows:

$$\frac{\left\| SD_k^{(t)} - SD_k^{(t-1)} \right\|}{\left\| SD_k^{(t-1)} \right\|} < 10^{-6} \quad (6)$$

where t denotes the number of iterations performed. Matrices $SM_{MSS}^{(t)}$, $SM_{MGIP}^{(t)}$ and $SM_{MSIE}^{(t)}$ represent the state matrices after t iterations, corresponding to states SM_{MSS}, SM_{MGIP}, and SM_{MSIE}, respectively.

The integrated network SD for the final three types of similarity networks is expressed as follows:

$$SD = \frac{SD_{DSS}^{(t)} + SD_{DGIP}^{(t)} + SD_{DSIE}^{(t)}}{3} \quad (7)$$

After the iteration ends, in order to convert the obtained matrix into a symmetric matrix, we use the following method to obtain the final fused similarity matrix:

$$SD' = \frac{SD + SD^T}{2} \quad (8)$$

The final fused similarity matrix for metabolites can be obtained using the same method.

3) Construct Heterogeneous Networks: To obtain richer graph information, we integrated metabolite-disease associations, disease-disease similarity information, and metabolite-metabolite similarity information, constructing a comprehensive network of metabolite-disease associations:

$$M(i,j) = \begin{cases} 0, & M_S(i,j) < S_m \\ 1, & M_S(i,j) \geq S_m \end{cases} \quad (9)$$

$$D(i,j) = \begin{cases} 0, & D_S(i,j) < S_d \\ 1, & D_S(i,j) \geq S_d \end{cases} \quad (10)$$

where $M(i,j)$ is the metabolite-metabolite association graph, M_S is the fused metabolite similarity network, S_m is the threshold for determining the relevance between metabolite nodes. $D(i,j)$ is the disease-disease association graph, D_S is the fused disease similarity network, and S_d is the threshold for determining the relevance between disease nodes.

Based on the metabolite-metabolite association information, we initialized a 2262$\times$ 216 all-zero adjacency matrix A to indicate whether there is a relationship between diseases and metabolites. If metabolite m_i is associated with disease d_j, the value of $A(i,j)$ is set to 1. otherwise, all other unknown metabolite-disease pairs and unrelated metabolite-disease pairs are set to 0.

The construction method of heterogeneous network is as follows:

$$Q = \begin{pmatrix} M & A \\ A^T & D \end{pmatrix} \quad (11)$$

where A denotes the metabolite-disease association matrix, and A^T is the transpose of A.

B. Features Extraction Process

1) Extraction of Association Matrix Features for Metabolites and Diseases by NMF: Each element in the metabolite-disease association matrix represents the degree of association between a specific metabolite and a disease, providing a visual representation of known associations between metabolites and diseases. This matrix serves as a rich source of input information for predictive models. NMF, or non-negative matrix factorization, is a technique that decomposes a large non-negative matrix into two non-negative matrices, such that the product of the decomposed matrices approximates the original matrix. It is an effective dimensionality reduction technique. In DCMDA, we apply NMF to decompose the metabolite-disease association matrix into a metabolite feature matrix and a disease feature matrix, aiming to achieve feature extraction.

$$A_{(m*d)} = M_{(m*k)} * D_{(k*d)} \quad (12)$$

where A represents the metabolite-disease association matrix, with m ($m = 2262$) denoting the number of metabolites, d ($d = 216$) representing the number of diseases, and k as the number of features extracted from A. M serves as the basis matrix in NMF, representing the extracted metabolite feature matrix, while D serves as the coefficient matrix in NMF, representing the extracted disease feature matrix.

To ensure that the product of the decomposed matrices closely approximates the original matrix, we formulated the following loss function. The loss function comprises a loss term and a regularization term. The loss term measures the discrepancy between the model's predicted values and the true values, while the

regularization term constrains the size of the model parameters to prevent overfitting to the training data. Our objective is to minimize the loss function to find appropriate M and D, thus approximating A.

$$S = \|W \odot (A - MD)\|_F^2 + \lambda_1 \|M\|_F^2 + \lambda_2 \|D\|_F^2 \quad (13)$$

where W is a matrix similar to A, while M and D are the decomposed matrices, sized 2262*90 and 90*216 respectively. The formula $\|.\|_F^2$ represents the Frobenius norm of the matrix, indicating the square root of the sum of squares of all elements in the matrix. The symbol $\odot$ denotes element-wise multiplication between two matrices. The symbol λ_1 and λ_2 denote regularization coefficients, After conducting hyperparameter tuning experiments, we set both λ_1 and λ_2 to 0.01. where the constraints on M and D necessitate them to be non-negative matrices.

Our objective is to minimize loss function S, which can be regarded as an optimization problem. We formulated the Lagrangian function of this optimization problem using the method of Lagrange multipliers, incorporating the loss term, regularization term, and Lagrange multipliers:

$$\begin{aligned} J(M, D) = & \|W \odot (A - MD)\|_F^2 + \lambda_1 T_r \left(MM^T\right) \\ & + \lambda_2 T_r \left(DD^T\right) + T_r \left(\varphi_{ik} \mathrm{M}^T\right) + T_r \left(\phi_{ki} \mathrm{D}^T\right) \end{aligned} \quad (14)$$

where Tr denotes the trace of a matrix, which is the sum of the elements on the matrix diagonal. φ_{ik} and φ_{ik} represent Lagrange multipliers. We aim to minimize the Lagrangian function to solve for matrices M and D.

Therefore, we take partial derivatives with respect to M and D, respectively.

$$\begin{aligned} \frac{\partial J}{\partial M} = & -2\left((W \odot A) D^T\right) + 2\left((W \odot (MD)) D^T\right) \\ & + 2\lambda_1 M + \varphi_{ik} \end{aligned} \quad (15)$$

$$\begin{aligned} \frac{\partial J}{\partial D} = & -2\left((W \odot A) M^T\right) + 2\left((W \odot (MD)) M^T\right) \\ & + 2\lambda_2 D + \phi_{ki} \end{aligned} \quad (16)$$

Based on the results of the partial derivatives, update rules were derived for iteratively updating matrices M and D:

$$m_{ik}^{(t)} = m_{ik}^{(t-1)} \frac{\left((W \odot A) D^T\right)_{ik}}{\left((W \odot (MD)) D^T + \lambda_1 M\right)_{ik}} \quad (17)$$

$$d_{ki}^{(t)} = d_{ki}^{(t-1)} \frac{\left((W \odot A) M^T\right)_{ki}}{\left((W \odot (MD)) M^T + \lambda_2 D\right)_{ki}} \quad (18)$$

Employing the aforementioned iterative method for 1000 updates, we ultimately obtained approximate matrices M and D. Matrix M represents the metabolite feature matrix, while the transpose of matrix D represents the disease feature matrix.

2) Extraction of Heterogeneous Network Features for Metabolites and Diseases by GATE: Multiscale feature extraction involves analyzing data at different scales to capture features across varying levels. This approach provides richer and more layered feature representations, thereby enhancing model performance and accuracy [56], [57], [58]. In addition to metabolite-disease association features, metabolite-metabolite association features and disease-disease association features also play a crucial role in predicting unknown metabolite-disease associations. Therefore, we input the heterogeneous graph containing all association features obtained from the data preprocessing stage into the GATE module for learning, aiming to extract more abundant association features between metabolites and diseases. The graph autoencoder [53] consists of an encoder and a decoder. The encoder is responsible for transforming input graph data into low-dimensional vector representations, while the decoder transforms the encoded low-dimensional vectors back into the original graph data. Graph data typically possesses complex structures and topological relationships, where connections between nodes can be quite intricate. In practical training, the model might struggle to capture these complex relationships between nodes, resulting in poor training performance. By introducing an attention mechanism [54], [55] to dynamically compute the correlation weights between nodes, the representation capability of node features can be enhanced. This enables the model to better capture the graph structure and learn the complex relationships between nodes, thus improving the performance and expressive power of the graph autoencoder model.

The attention weights are dynamically computed based on the similarity between node features, thereby weighting the importance of different nodes. The attention mechanism generates a sparse attention weight matrix, which is used to combine node features with weights, thereby obtaining the output of the encoder.

$$\begin{cases} H_o = Softmax\left(X\left((QW_o)V_{|0|}\right) + X\left((QW_o)V_{|1|}\right)^T\right) * (QW_o) \\ H_{i+1} = Softmax\left(X\left((H_iW_{i+1})V_{|0|}\right) + X\left((H_iW_{i+1})V_{|1|}\right)^T\right) \\ \qquad * (H_iW_{i+1}), 0 <= i <= \max \end{cases} \quad (19)$$

In each layer of the decoder, the same attention weight matrix as the encoder is utilized to reconstruct the encoder output into the original node features.

$$\begin{cases} H'_{i+1} = Softmax\left(X\left((H_iW_{i+1})\,V_{|0|}\right) + X\left((H_iW_{i+1})\,V_{|1|}\right)^T\right) \\ \qquad * (H_{i+1}W_{i+1}), i = \max \\ H'_{i+1} = Softmax\left(X\left((H_iW_{i+1})\,V_{|0|}\right) + X\left((H_iW_{i+1})\,V_{|1|}\right)^T\right) \\ \qquad * \left(H'_{i+2}W_{i+1}\right), 0 <= i < \max \\ H'_o = Softmax\left(X\left((QW_o)\,V_{|0|}\right) + X\left((QW_o)\,V_{|1|}\right)^T\right) \\ \qquad * (H'_1W_o) \end{cases} \quad (20)$$

where Q and X are different representations of the complete disease-metabolite association network. W represents the trainable weight matrices of different layers of the encoder and decoder, while V represents the trainable parameters. Softmax() is used to normalize the attention weight matrix.

The training objective of the graph autoencoder is to make the output of the decoder as close as possible to the original graph data. To achieve this goal, DCMDA's loss function needs to be computed, consisting of two parts: structural reconstruction loss

and feature reconstruction loss. The feature reconstruction loss measures the difference between the output of the decoder and the original input features, and it can be expressed as:

$$features\_loss = \|Q - H'\|_F^2 \tag{21}$$

where Q represents the original input features of the encoder, H' is the output features of the decoder, $\|.\|_F^2$ denotes the Frobenius norm of a matrix, representing the square root of the sum of squares of all elements in the matrix.

The structural reconstruction loss measures the reconstruction error of the graph structure, specifically the error in reconstructing the relationships between nodes. It can be represented as follows:

$$structure\_loss = \sum_{j=1}^{P} St\left(S_j, R_j\right) \tag{22}$$

$St(S, R)$represents the similarity loss between node S and node R, where p denotes the number of node pairs. The total structural reconstruction loss is obtained by summing the similarity losses between all pairs of nodes.

The calculation formula for $St(S, R)$ is as follows:

$$St(S, R) = -\log\left(sigmoid\left(\sum_{i=1}^{\dim}\left(H_{Si} * H_{Ri}\right)\right)\right) \tag{23}$$

where H_S and H_R respectively denote the output feature vectors from the encoders corresponding to nodes S and R. dim represents the dimensionality of the feature vectors. The similarity between two nodes is computed by element-wise multiplication of their feature vectors followed by summation of all elements in the resulting vector. The $sigmoid()$ function transforms the sum to a probability within the range [0, 1]. The $log()$ function takes the natural logarithm of the transformed value. Finally, by taking the negative value, the cosine similarity is converted to similarity loss, meaning that as the similarity between node pairs increases, the similarity loss decreases.

Finally, by combining the feature reconstruction loss and the structural reconstruction loss with weighted factors, DCMDA's parameters are optimized by minimizing the total loss. Here, λ serves as a hyperparameter that controls the weight of the structural reconstruction loss.

$$loss = features\_loss + \lambda * structure\_loss \tag{24}$$

C. Feature Fusion Module

For the association matrix features and heterogeneous network features extracted by the feature extraction module, we utilize linear transformations to reduce their dimensions to the same level. Then, the two sets of reduced features undergo Cross-Attention mechanisms to obtain a fused feature representation [51], [52]. This fused representation is further transformed through linear transformations, and after passing through an activation function, the final metabolite-disease association score is derived.

Employing the Cross-Attention mechanism effectively integrates two related but distinct metabolite-disease features. In the feature extraction module, DCMDA utilizes NMF to extract association matrix features of metabolites and diseases from the metabolite-disease association network. Additionally, it leverages GATE to achieve feature learning across the entire metabolite-disease association network (including relationships between diseases, relationships between metabolites, and relationships between diseases and metabolites), thus obtaining heterogeneous network features of metabolites and diseases. Since the association matrix features and heterogeneous network features possess their respective specificities, we can integrate these two types of features through the Cross-Attention mechanism, enabling the heterogeneous network features to focus on the key information of the association matrix features, ultimately achieving feature fusion and obtaining a deep representation of disease-metabolite interactions.

To effectively utilize the Cross-Attention mechanism for feature fusion, we make the heterogeneous network features attend to the association matrix features. We take the association matrix features output by NMF as input X and the heterogeneous network features output by GATE as input Y for cross-attention computation. Initially, we compute the query matrix Q, key matrix K, and value matrix V for each input feature x and y. The expressions are as follows:

$$Q_y = YW^Q \tag{25}$$

$$K_x = XW^k \tag{26}$$

$$V_x = XW^v \tag{27}$$

The query matrix Q is multiplied with the key matrix K, and a softmax normalization is applied to obtain a normalized attention weight matrix. Finally, the Cross-Attention output is obtained through the dot product operation between the attention weight matrix and the value matrix V, expressed as:

$$cross\_Attention(Q_y,K_x,V_x) = Soft\max\left(\frac{Q_y \cdot K_x^T}{\sqrt{dk}}\right) \cdot V_x \tag{28}$$

where $\sqrt{dk}$ represents the dimensionality of the key matrix to prevent excessively large computations.

We can observe that the heterogeneous network features dynamically attend to the primary features of the association matrix, effectively combining key information from both parts and fusing the feature information of metabolites and diseases.

After obtaining the fused features of metabolites and diseases through the Cross-Attention mechanism, we use the binary cross-entropy function for the final classification. Binary cross-entropy is a loss function used for binary classification problems, measuring the uncertainty between the predicted values and the actual labels.

$$\text{BC} = -\left(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right) \tag{29}$$

In the equation above, the actual label y takes values of either 0 or 1, and the predicted value $\hat{y}$ represents a probability between 0 and 1. When the actual label $y = 1$, the loss function is $-\log(\hat{y})$, which indicates that the closer the predicted probability $\hat{y}$ is to 1, the smaller the loss. When the actual label $y = 0$, the loss

function is $-\log(1-\hat{y})$, indicating that the closer the predicted probability $\hat{y}$ is to 0, the smaller the loss.

III. Result

A. Experimental Environment Setup

Our experiment was carried out on a server equipped with NVIDIA Geforce RTX A6000 GPU and Intel Core i9-13900K CPU running on Linux (Ubuntu 20.04) operating system. For NMF, the value of k is set to 90, with regularization coefficients $\lambda_1 = \lambda_2 = 0.01$, and 1000 iterations are performed, resulting in $M_{2262*90}$ and D_{90*216}. For GATE, the number of neurons in the encoder layers are set to (128,64), and in the decoder layers to (64,128). GATE undergoes unsupervised learning for 300 epochs. The Cross-Attention module comprises 2 hidden layers, with a batch size of 128, utilizing the Adam optimizer, and trains for 100 epochs to obtain the optimal weights. Additional details of hyperparameter selection process are presented in results section.

B. Cross-Validation and Experimental Metrics

To validate the effectiveness of DCMDA, we conducted five-fold cross-validation. The dataset was divided into five equally sized subsets, with one subset reserved as the validation set and the remaining four subsets used for training. Subsequently, DCMDA was trained on these four training subsets and evaluated on the reserved validation subset. This process was repeated five times, each time using a different validation subset. Ultimately, the evaluation of DCMDA's performance is the average of the results from these five validations. Five-fold cross-validation allows for assessing DCMDA's generalization ability, testing its stability, and reducing biases introduced by improper data splitting. By repeating cross-validation multiple times and averaging the results, we can more reliably assess DCMDA's performance on different subsets of data, thereby gaining a better understanding of its overall performance.

To build the predictive model, DCMDA is trained on known associations in the training set and then used to forecast associations in the test set to obtain corresponding results. Subsequently, the true positive rate (TPR) and false positive rate (FPR) can be calculated using the following formulas:

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

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

In the above equation, TP and TN respectively represent the number of correctly predicted positive and negative samples. FN and FP represent the number of falsely predicted negative and positive samples. The main experimental metrics for evaluating DCMDA are AUC and AUPR. AUC represents the area under the ROC (Receiver Operating Characteristic) curve, which plots the relationship between the true positive rate (TPR) and the false positive rate (FPR). The AUC value ranges between 0 and 1, where a value closer to 1 indicates better classification ability of the DCMDA, while a value closer to 0.5 indicates performance close to random classification. In AUPR, the horizontal axis typically represents $recall$, while the vertical axis represents $precision$. The $precision - recall$ curve primarily focuses on the trade-off between $precision$ and $recall$ of positive samples. A higher AUPR value indicates a better balance between $precision$ and $recall$. $Precision$ represents the number of true positive samples among those predicted as positive, while $recall$ refers to the proportion of actual positive samples predicted as positive. The calculation formulas are as follows:

$$Precision(P) = \frac{TP}{TP + FP} \tag{32}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{33}$$

For vertical experimental comparisons, we included additional metrics. $Accuracy$ is the proportion of correctly classified samples to the total number of samples in the given test dataset, representing the probability of correct predictions by

$$Acccuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{34}$$

The F1-Score is the weighted harmonic mean of $precision$ and $recall$ by

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \tag{35}$$

C. Comparing With Other Binary Classification Models

Currently, numerous researchers have proposed various association prediction methods for application in the field of bioinformatics. In this study, we have selected five representative association prediction methods for comparison with DCMDA and tested each of them on the same dataset to highlight the advantages of our approach. These five methods are MGDHGS [12], MAHN [19], MDA-AENMF [13], EKRR [20], and GCNAT [16]. Below, we provide a brief introduction to these five methods:

1) MGDHGS employs GraphSAGE to sample and aggregate features from the local neighborhoods of nodes to generate embeddings. Additionally, it utilizes a self-attention mechanism to achieve adaptive allocation of a large number of weights between metabolites and diseases.
2) MAHN utilizes GCN and enhanced GraphSAGE to mine semantic network information with a meta-path length of 3. It also employs node-level and semantic-level attention to explore deeper and more complex features from semantic networks with a meta-path length of 2.
3) MDA-AENMF extracts features from three different modules, then combines these features into a single composite feature vector for each metabolite-disease pair. Finally, the corresponding feature vectors and labels are fed into a multilayer perceptron for training.
4) EKRR employs an ensemble learning and feature dimension reduction computational framework. It utilizes random selection of features to create multiple base classifiers by combining two Kernel Ridge Regression classifiers—one for the miRNA side and the other for the disease side.

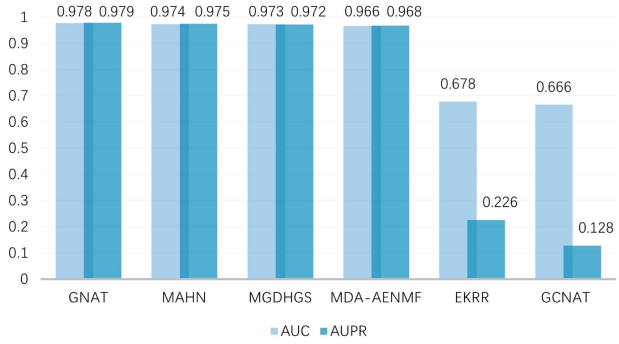

Fig. 2. AUC and AUPR of DCMDA and comparison methods by 5-fold CV under the same dataset.

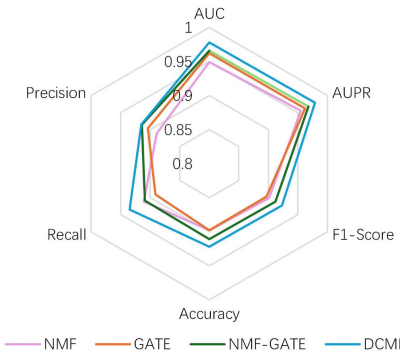

Fig. 3. Comparison analysis between DCMDA and its ablation experiments.

Subsequently, an averaging strategy is employed on these base classifiers to obtain the final association scores for miRNA-disease pairs.

5) In GCNAT, encoding and learning are performed based on graph convolutional neural networks (GCNs) to predict potential associations between disease-related metabolites.

To comprehensively evaluate the predictive performance of DCMDA and other comparative models, we conducted separate 5-fold cross-validation experiments on the same dataset. As shown in Fig. 2, DCMDA achieved the highest scores in both AUC and AUPR, with values of 0.978 and 0.979, respectively. In contrast, the lower score of EKRR may be attributed to its reliance on optimization and complex network algorithms, without incorporating neural networks and attention mechanisms, leading to inferior performance. Although MGDHGS, MAHN, and MDA-AENMF also utilized neural networks, they extracted features by separately inputting metabolite and disease information into the neural network, without considering the latent features between metabolites and diseases. While GCNAT fed the fused network of metabolites and diseases into the neural network for simultaneous feature extraction, the fused network consisted of only three similarity networks: disease Gaussian kernel similarity, disease semantic similarity, and metabolite Gaussian kernel similarity, which provided insufficient information. DCMDA, on the other hand, generated a fused heterogeneous network by integrating seven similarity networks, resulting in richer network information being input into the neural network for feature extraction. Additionally, we employed non-negative matrix factorization to extract another set of association features between metabolites and diseases. Furthermore, considering the specificity of the information extracted by different modules, we adopted a cross-attention mechanism to fuse this information, automatically adjusting the importance of different features and suppressing irrelevant ones. This allows DCMDA to focus on more useful information, obtain deep representations of metabolites and diseases, and ultimately achieve better prediction results.

D. Ablation Study

Since Cross-Attention is a mechanism that requires blending two different embedding sequences, it is not feasible to retain the NMF+Cross-Attention group or the GATE+Cross-Attention group in the ablation experiments. To demonstrate the effectiveness of each component, we adopted a strategy of progressively adding components, divided into four groups:

1) NMF: Only the five-layer autoencoder.
2) GATE: Only graph autoencoders equipped with attention mechanisms.
3) NMF-GATE: The combination of a five-layer autoencoder and a graph autoencoder equipped with an attention mechanism.
4) DCMDA: The combination of a five-layer autoencoder, a graph autoencoder equipped with an attention mechanism, and a Cross-Attention mechanism.

As shown in Fig. 3, compared to Group NMF, Group NMF-GATE saw increases of 1.8%, 1.4%, 1.1%, 1.3%, and 2.5% in AUC, AUPR, F1-Score, Accuracy, and Precision, respectively, while experiencing a decrease of 0.2% in Recall. Compared to Group GATE, Group NMF-GATE observed increases of 0.5%, 0.7%, 1.5%, 1.3%, 1.8%, and 0.1% in AUC, AUPR, F1-Score, Accuracy, Recall, and Precision, respectively. The experimental results indicate that the predictive performance of Group NMF-GATE, which combines the concatenated NMF output features and GATE output features, surpasses that of Group NMF, which solely employs NMF features, and Group GATE, which solely utilizes GATE features, thus demonstrating the effectiveness of integrating both NMF and GATE components. Furthermore, by utilizing the Cross-Attention mechanism to combine the output features of NMF and GATE in Group DCMDA, the optimal performance is achieved. Compared to Group NMF-GATE, increases of 1.2%, 1.1%, 1.1%, 1.2%, 2.6%, and 0.1% were observed in AUC, AUPR, F1-Score, Accuracy, Recall, and Precision, respectively. This is attributed to the enhanced focus on critical information achieved by integrating these distinct features, leading to more profound representations of metabolites and disease characteristics. In summary, the combination of the feature extraction module NMF and GATE, along with the feature fusion module utilizing Cross-Attention, is indispensable, as they play crucial roles in predicting potential associations between metabolites and diseases.

E. Comparison Between Cross-Attention and Self-Attention

In this experiment, we conducted a comparative study between Cross-Attention and Self-Attention. Cross-Attention combines two embedding sequences of the same dimension

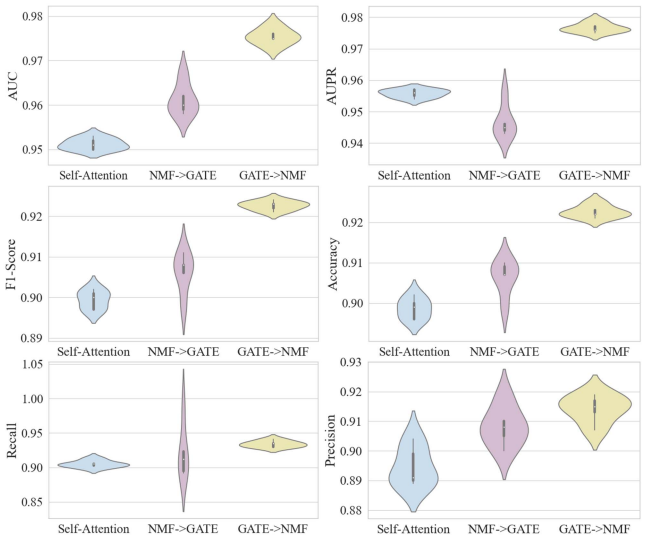

Fig. 4. Comparison analysis between Cross-Attention and Self- Attention.

asymmetrically, where one sequence serves as the query Q input, while the other serves as the key K and value V inputs. In contrast, Self-Attention takes a single embedding sequence as input. When using Cross-Attention, we treated the outputs of NMF and GATE as two separate embedding sequences. Additionally, we explored which of the outputs of NMF or GATE would yield better results as the query Q input and conducted experiments accordingly. For the Self-Attention mechanism, we first concatenated the outputs of NMF and GATE and then inputted them into the Self-Attention mechanism.

As depicted in Fig. 4, compared to the Self-Attention group, the Cross-Attention group where NMF focuses on GATE achieved increases of 1.1%, 1.1%, 1.1%, 0.7%, and 1.9% in AUC, F1-Score, Accuracy, Recall, and Precision, respectively, while experiencing a decrease of 1% in AUPR. In comparison to the Cross-Attention group where NMF focuses on GATE, the Cross-Attention group where GATE focuses on NMF witnessed increases of 1.6%, 3.2%, 1.5%, 1.5%, 2.3%, and 0.7% in AUC, AUPR, F1-Score, Accuracy, Recall, and Precision, respectively. The experimental results indicate that the Cross-Attention group where GATE focuses on NMF achieved the highest scores, with all six metrics outperforming the other two groups. This outcome demonstrates the effectiveness of the GATE output features attending to NMF output features within the Cross-Attentionn mechanism, validating the efficient integration of the components in this model.

F. External Validation

To rigorously validate DCMDA's excellent generalization performance, we employed the dataset from Deep-DRM [21], which also originates from HMDB and includes 1,436 metabolites, 242 diseases, and 3,124 known associations between diseases and metabolites. This dataset is smaller than the one used in this study. We conducted comparative experiments between Deep-DRM, MDA-AENMF [18], and DCMDA. As shown in Fig. 5, the results indicate that Deep-DRM achieved AUC and

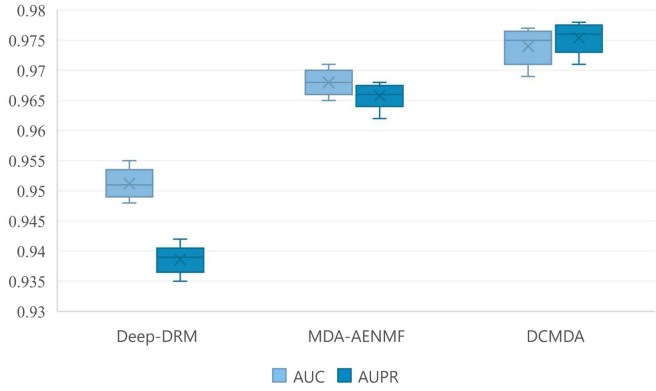

Fig. 5. Comparison results between DCMDA, MDA-AENMF and Deep-DRM.

TABLE I
COMPARISON ANALYSIS OF DIFFERENT CROSS-ATTENTION HIDDEN LAYERS

Cross-Attention hidden layers	AUC	AUPR
128	0.970	0.971
128 64	0.978	0.979
128 64 32	0.971	0.972
128 64 32 16	0.959	0.913

TABLE II
COMPARISON ANALYSIS OF DIFFERENT GATE HIDDEN LAYERS

GATE hidden layers	AUC	AUPR
128	0.963	0.973
128 64	0.978	0.979
128 64 32	0.949	0.966
128 64 32 16	0.933	0.959

AUPR values of 0.952 and 0.939, respectively. MDA-AENMF outperformed Deep-DRM by 1.7% in AUC and 2.7% in AUPR, while DCMDA outperformed Deep-DRM by 2.3% in AUC and 3.8% in AUPR. These experimental results suggest that DCMDA exhibits superior generalization performance compared to Deep-DRM and MDA-AENMF, demonstrating faster adaptation to new data features and distributions and robust applicability across various practical scenarios.

G. Parameter Analysis and Optimization Experiment

As shown in Table I, when the number of layers in the feature fusion module is set to 1, the model's training performance is slightly lower, with AUC and AUPR scores of 0.97 and 0.971, respectively. When the hidden layer is increased to 2 layers, the model's performance improves. Compared to a single hidden layer, AUC and AUPR increase by 0.8% each. However, further increasing the number of hidden layers to 3 or 4 does not result in significant improvements in training performance compared to the model with 2 hidden layers. Additionally, training time increases. Considering both time cost and performance, we set the number of stacked layers in the feature fusion module to two, balancing efficient computation with satisfactory prediction accuracy. The impact of GATE's hidden layers on model training performance is similar to that of the Cross-Attention mechanism, as shown in Table II, with the best results achieved when the hidden layers are set to two. Regarding the regularization

TABLE III
COMPARISON ANALYSIS OF DIFFERENT λ_1 AND λ_2

λ_1 and λ_2	AUC	AUPR
0.001	0.945	0.963
0.005	0.949	0.971
0.01	0.978	0.979
0.02	0.958	0.975
0.03	0.956	0.973

TABLE IV
TOP 10 POTENTIAL METABOLITES ASSOCIATED WITH LEUKEMIA

Leukemia			
Rank	Metabolite name	Evidences	Confirmed
1	Creatinine	HMDB0000562	Yes
2	L-Lactic acid	HMDB0000190	Yes
3	D-Glucose	HMDB0000122	No
4	L-Tyrosine	HMDB0000158	Yes
5	L-Phenylalanine	HMDB0000159	Yes
6	L-Serine	HMDB0000187	Yes
7	Glycine	HMDB0000123	Yes
8	Ornithine	HMDB0000214	Yes
9	Biotin	HMDB0000030	Yes
10	Pyruvic	HMDB0000243	Yes

coefficients λ_1 and λ_2 for non-negative matrix factorization, as presented in Table III, the optimal model training performance occurs when $\lambda_1 = \lambda_2 = 0.01$.

H. Case Study

To provide a more intuitive understanding of DCMDA's performance in identifying potential metabolite-disease associations, we conducted case studies on four common diseases: leukemia, uremia, obesity, and hepatitis. For each disease, we predicted the scores of all metabolites in the dataset and ranked them in descending order. We then selected the top ten associations for analysis.

Leukemia, a common malignancy in both children and adults, arises when alterations in normal cellular regulatory processes lead to the uncontrolled proliferation of hematopoietic stem cells in the bone marrow. Symptoms of leukemia are nonspecific and include fever, fatigue, weight loss, bone pain, bruising, and bleeding [22]. In the case study of leukemia, we examined the top 10 associations, with validation results shown in Table IV. Among them, L-Tyrosine, L-Phenylalanine, L-Serine, Glycine, and Glycine scored highly, and their associations with leukemia were validated in 4,536 metabolite-disease pairs from HMDB in this study. Although Creatinine, L-Lactic Acid, Biotin, and Pyruvic Acid had no direct validation evidence from the dataset, we found supporting evidence in previous studies linking these metabolites to leukemia [23], [34], [35], [36]. For example, creatinine excretion rates were consistently lower in leukemia cases [23], and upregulation of phosphoserine phosphatase (PSPH) was frequently observed in T-cell acute lymphoblastic leukemia, correlating with elevated serine and glycine levels in xenotransplantation mouse models [24]. D-Glucose, ranked third by score, did not have validation evidence in this dataset, other datasets, or existing literature linking it to leukemia.

The characteristic feature of uremia is the retention of various compounds that, in a healthy individual, are excreted by the kidneys into the urine. These compounds interfere with many physiological functions, leading to toxicity [25]. In the uremia case studies shown in Table V, the association of 10 metabolites with uremia was confirmed. For Hippuric acid, ranked seventh, and Guanidoacetic acid, ranked ninth, we found evidence of their association with uremia in the original dataset. Additionally, for the remaining eight metabolites among the top 10 associated with uremia, evidence of their associations can be found in previous studies [37], [38], [39], [40], [41]. For example, amino acid analysis of the liver in uremic animals showed increased concentrations of L-Alanine [26]. In comparison to the control group, uremic patients exhibited lower levels of all cationic amino acids (L-Arginine, L-Ornithine, and L-Lysine) in plasma [27]. Uremic patients lack functional renal tissue, and the primary source of endogenous L-Arginine for total NO synthase is from the normal renal cortex [28].

TABLE V
TOP 10 POTENTIAL METABOLITES ASSOCIATED WITH UREMIA

Uremia			
Rank	Metabolite name	Evidences	Confirmed
1	L-Alanine	HMDB0000161	Yes
2	L-Tyrosine acid	HMDB0000158	Yes
3	Creatine	HMDB0000064	Yes
4	L-Lysine	HMDB0000182	Yes
5	L-Tryptophan	HMDB0000929	Yes
6	L-Phenylalanine	HMDB0000159	Yes
7	Hippuric acid	HMDB0000714	Yes
8	L-Arginine	HMDB0000517	Yes
9	Guanidoacetic acid	HMDB0000128	Yes
10	L-Histidine	HMDB0000177	Yes

Obesity is a complex disease influenced by numerous causal factors, many of which are largely beyond individual control. It leads to significant suffering, poor health, impaired function, reduced quality of life, serious illnesses, and increased mortality rates [29]. We conducted a case study on obesity, examining newly predicted associations between obesity and metabolites ranked in the top 10 based on prediction scores, as shown in Table VI. Nine of the top ten metabolites had confirmed associations with obesity in the original dataset. However, for 3-Hydroxy-cis-5-tetradecenoylcarnitine, no association with obesity was found in the original dataset, nor could it be confirmed in previous research.

Hepatitis is a general term for liver inflammation, which can be caused by various viruses [30]. Clinical symptoms of hepatitis include jaundice, fatigue, nausea, and loss of appetite [31]. In the hepatitis case study, we examined the top 10 predicted associations, none of which were found in the original dataset. However, nine out of these ten associations were validated in previous studies [42], [43], [44], [45], [46], [47], [48], as shown in Table VII. For instance, research suggests that cholesterol metabolism is crucial for the hepatitis virus infection and its lifecycle [32]. Additionally, neuropeptide Y and substance P, released by nerve fibers and immune cells, contribute to hepatitis inflammation and the resolution of inflammation [33]. However, the association between D-Glucose and hepatitis has not yet been validated.

TABLE VI
TOP 10 POTENTIAL METABOLITES ASSOCIATED WITH OBESITY

Obesity			
Rank	Metabolite name	Evidences	Confirmed
1	Linoleyl carnitin	HMDB0006469	Yes
2	Lycopene	HMDB0003000	Yes
3	PC(18:3(6Z,9Z,12Z)/P-18:0)	HMDB0008193	Yes
4	PC(18:1(9Z)/18:3(6Z,9Z,12Z))	HMDB0008106	Yes
5	PC(18:3(9Z,12Z,15Z)/P-18:1(11Z))	HMDB0008227	Yes
6	PC(18:1(9Z)/20:3(5Z,8Z,11Z))	HMDB0008112	Yes
7	3-Hydroxy-cis-5-tetradecenoylcarnitine	HMDB0013330	Yes
8	PC(18:0/18:2(9Z,12Z))	HMDB0008039	Yes
9	PC(18:3(6Z,9Z,12Z)/18:0)	HMDB0008168	Yes
10	PC(18:1(11Z)/18:3(9Z,12Z,15Z))	HMDB0008074	Yes

TABLE VII
TOP 10 POTENTIAL METABOLITES ASSOCIATED WITH HEPATITIS

Hepatitis			
Rank	Metabolite name	Evidences	Confirmed
1	Cholesterol	HMDB0000067	Yes
2	Homocysteine acid	HMDB0000742	Yes
3	Homovanillic acid	HMDB0000118	Yes
4	Phosphate	HMDB0001429	Yes
5	Uric acid	HMDB0000289	Yes
6	Epinephrine	HMDB0000068	Yes
7	D-Glucose	HMDB0000122	No
8	Indoleacetic acid	HMDB0000197	Yes
9	Substance P acid	HMDB0001897	Yes
10	Potassium	HMDB0000586	Yes

IV. CONCLUSION

In this study, we propose a metabolite-disease association prediction framework called DCMDA, which integrates nonlinear methods to combine distinct similarity networks of metabolites and diseases. It learns from these fused similarity networks along with the metabolite-disease association network. First, two sets of features one for metabolites and one for diseases are separately extracted from the raw data using the NMF and GATE modules. Next, low-dimensional representations of these extracted features are generated using linear layers. The final features, obtained by fusing the two specific low-dimensional feature representations via Cross-Attention, are processed through dimensionality reduction and normalized using a sigmoid function to produce the final association scores for metabolite-disease pairs. We demonstrate the effectiveness of DCMDA through experiments, where it outperforms state-of-the-art methods with an AUC score of 0.978 and an AUPR score of 0.979. The robustness and reliability of DCMDA are further validated through generalization studies on additional datasets, where it achieves significantly higher scores than comparative models, demonstrating its excellent generalization ability and adaptability. Moreover, case studies on four common diseases: leukemia, obesity, uremia, and hepatitis, show DCMDA's high accuracy in predicting potential metabolite-disease associations, further confirming its effectiveness.

However, DCMDA still faces some limitations and challenges. Due to the limited amount of data and the vast diversity of metabolites and diseases, the model is not applicable to new metabolites and diseases. Nevertheless, compared to previous methods, DCMDA exhibits strong performance in predicting metabolite-disease associations. In the future, the cold start problem could be addressed by incorporating new contrastive learning techniques, further improving the model's generalization on new datasets.

REFERENCES

[1] W. B. Dunn and D. I. Ellis, "Metabolomics: Current analytical platforms and methodologies," *TrAC Trends Anal. Chem.*, vol. 24, no. 4, pp. 285–294, 2005.

[2] L. Cheng et al., "MetSigDis: A manually curated resource for the metabolic siDCMDAures of diseases," *Brief. Bioinf.*, vol. 20, no. 1, pp. 203–209, 2019.

[3] J. Lu et al., "Metabolomics in human type 2 diabetes research," *Front. Med.*, vol. 7, no. 1, pp. 4–13, 2013.

[4] Y.-M. Chen et al., "Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-alcoholic fatty liver disease in adults," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 19076.

[5] S. Chen et al., "*Fusobacterium nucleatum* promotes colorectal cancer metastasis by modulating *KRT7-AS*/KRT7," *Gut Microbes*, vol. 11, no. 3, pp. 511–525, 2020.

[6] Y. Heianza et al., "Gut microbiota metabolites and risk of major adverse cardiovascular disease events and death: A systematic review and meta-analysis of prospective studies," *J. Amer. Heart Assoc.*, vol. 6, no. 7, 2017, Art. no. 4947.

[7] C. Astore and G. Gibson, "Integrative polygenic analysis of the protective effects of fatty acid metabolism on disease as modified by obesity," *Front. Nutr.*, vol. 10, no. 1, 2024, Art. no. 1308622.

[8] L. E. Louck et al., "The relationship of circulating choline and choline-related metabolite levels with health outcomes: A scoping review of genome-wide association studies and Mendelian randomization studies," *Adv. Nutr.*, vol. 1, no. 1, 2023, Art. no. 100164.

[9] S. Hori et al., "A metabolomic approach to lung cancer," *Lung Cancer*, vol. 74, no. 2, pp. 284–292, 2011.

[10] C. Czech et al., "Metabolite profiling of alzheimer's disease cerebrospinal fluid," *PLoS One*, vol. 7, no. 2, 2012, Art. no. 31501.

[11] S. Bhattacharya et al., "Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery disease in patients referred for cardiac catheterization," *Atherosclerosis*, vol. 232, no. 1, pp. 191–196, 2014.

[12] Y. Hu et al., "Identifying diseases-related metabolites using random walk," *BMC Bioinf.*, vol. 19, no. 1, pp. 37–46, 2018.

[13] X. Lei and C. Zhang, "Predicting metabolite-disease associations based on katz model," *BioData Mining*, vol. 12, no. 1, pp. 1–14, 2019.

[14] X. Lei, J. Tie, and H. Fujita, "Relational completion based non-negative matrix factorization for predicting metabolite-disease associations," *Knowl.-Based Syst.*, vol. 204, no. 1, 2020, Art. no. 106238.

[15] C. Zhang, X. Lei, and L. Liu, "Predicting metabolite–disease associations based on lightgbm model," *Front. Genet.*, vol. 12, no. 1, 2021, Art. no. 660275.

[16] F. Sun, J. Sun, and Q. Zhao, "A deep learning method for predicting metabolite–disease associations via graph neural network," *Brief. Bioinf.*, vol. 23, no. 4, 2022, Art. no. 266.

[17] J. Tie, X. Lei, and Y. Pan, "Metabolite-disease association prediction algorithm combining deepwalk and random forest," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 58–67, 2021.

[18] H. Gao et al., "Predicting metabolite–disease associations based on auto-encoder and non-negative matrix factorization," *Brief. Bioinf.*, vol. 24, no. 5, 2023, Art. no. 259.
[19] E. J. Yates and L. C. Dixon, "Pagerank as a method to rank biomedical literature by importance," *Source Code Biol. Med.*, vol. 10, no. 1, pp. 1–9, 2015.
[20] L.-H. Peng et al., "A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression," *Front. Bioeng. Biotechnol.*, vol. 8, no. 1, 2020, Art. no. 40.
[21] T. Zhao, Y. Hu, and L. Cheng, "Deep-DRM: A computational method for identifying disease-related metabolites based on graph deep learning approaches," *Brief. Bioinf.*, vol. 22, no. 4, 2021, Art. no. 212.
[22] A. S. Davis, A. J. Viera, and M. D. Mead, "Leukemia: An overview for primary care," *Amer. Fam. Physician*, vol. 89, no. 9, pp. 731–738, 2014.
[23] M. Atamer and A. Dietz, "Creatine and creatinine excretion in leukemia," *J. Lab. Clin. Med.*, vol. 58, no. 1, pp. 95–103, 1961.
[24] K. R. Kampen et al., "Translatome analysis reveals altered serine and glycine metabolism in T-cell acute lymphoblastic leukemia cells," *Nat. Commun.*, vol. 10, no. 1, 2019, Art. no. 2542.
[25] R. Vanholder et al., "What is uremia? Retention versus oxidation," *Blood Purif.*, vol. 24, no. 1, pp. 33–38, 2005.
[26] K. Maier et al., "Enzymatic and metabolic studies on carbohydrate and amino acid metabolism in rat liver during acute uraemia," *Eur. J. Clin. Investigation*, vol. 3, no. 3, pp. 201–207, 1973.
[27] P. F. Reis et al., "Plasma amino acid profile and l-arginine uptake in red blood cells from malnourished uremic patients," *J. Renal Nutr.*, vol. 16, no. 4, pp. 325–331, 2006.
[28] S. Xiao et al., "Uremic levels of urea inhibit L-arginine transport in cultured endothelial cells," *Amer. J. Physiol.-Renal Physiol.*, vol. 280, no. 6, pp. 989–995, 2001.
[29] D. B. Allison et al., "Obesity as a disease: A white paper on evidence and arguments commissioned by the council of the obesity society," *Obesity*, vol. 16, no. 6, 2008, Art. no. 1161.
[30] N. Previsani, D. Lavanchy, and G. Siegl, "Hepatitis A," *Perspectives Med. Virol.*, vol. 10, no. 1, pp. 1–30, 2003.
[31] N. McIntyre, "Clinical presentation of acute viral Hepatitis," *Brit. Med. Bull.*, vol. 46, no. 2, pp. 533–547, 1990.
[32] J.-W. Liou, H. Mani, and J.-H. Yen, "Viral Hepatitis, cholesterol metabolism, and cholesterol-lowering natural compounds," *Int. J. Mol. Sci.*, vol. 23, no. 7, 2022, Art. no. 3897.
[33] E. Fehér, "Changes in neuropeptide Y and substance P immunoreactive nerve fibres and immunocompetent cells in Hepatitis," *Orvosi Hetilap*, vol. 156, no. 47, pp. 1892–1897, 2015.
[34] Z.-W. Huang et al., "STAT5 promotes PD-l1 expression by facilitating histone lactylation to drive immunosuppression in acute myeloid leukemia," *Signal Transduct. Target. Ther.*, vol. 8, no. 1, 2023, Art. no. 391.
[35] K. Wróbel et al., "Cytarabine and dexamethasone-pamam dendrimer di-conjugate sensitizes human acute myeloid leukemia cells to apoptotic cell death," *J. Drug Del. Sci. Technol.*, vol. 81, no. 1, 2023, Art. no. 104242.
[36] C. P. Carretero et al., "TRAF3-inactivated chronic lymphocytic leukemia cells show an enhanced metabolic plasticity that can be attenuated by glutaminolysis and mitochondrial pyruvate import inhibition," *Blood*, vol. 142, no. 1, 2023, Art. no. 1889.
[37] D. Lubkowicz, "Engineered Escherichia coli Nissle 1917 for the prevention of uremic toxin accumulation in chronic kidney disease," Ph.D. dissertation, College Sci., Northeastern University, 2023.
[38] R. Hang, "Pectin in the reversal of uremia in renal failure," 2024.
[39] A. Tomášová et al., "The relationship of uremic toxin indoxyl sulfate and intestinal elimination mechanisms in hemodialysis patients," *Kidney Blood Press. Res.*, vol. 48, no. 1, pp. 28–34, 2023.
[40] S. Arefin et al., "Associations of biopterins and adma with vascular function in peripheral microcirculation from patients with chronic kidney disease," *Int. J. Mol. Sci.*, vol. 24, no. 6, 2023, Art. no. 5582.
[41] J.-K. Kim et al., "Metabolic and transcriptomic changes in the mouse brain in response to short-term high-fat metabolic stress," *Metabolites*, vol. 13, no. 3, 2023, Art. no. 407.
[42] A. Moghe, B. M. McGuire, and C. Levy, "Acute hepatic porphyrias–a guide for hepatologists," *Hepatology*, vol. 1, no. 1, 2024, Art. no. 1.
[43] T. Liu et al., "Bacillus coagulans regulates gut microbiota and ameliorates the alcoholic-associated liver disease in mice," *Front. Microbiol.*, vol. 15, no. 1, 2024, Art. no. 185.
[44] K. Murata et al., "Immunomodulatory mechanism of acyclic nucleoside phosphates in treatment of Hepatitis B virus infection," *Hepatology*, vol. 71, no. 5, pp. 1533–1545, 2020.
[45] P. Xu et al., "The relationship between serum uric acid level and liver function in patients with Hepatitis B in China," *Clin. Lab.*, vol. 1, no. 5, 2021, Art. no. 1105.
[46] B. S. Hofer et al., "Decreased platelet activation predicts hepatic decompensation and mortality in patients with cirrhosis," *Hepatology*, vol. 1, no. 1, 2023, Art. no. 1.
[47] J. Yamada, Y. Sugimoto, and K. Horisaka, "Elevation of serum indoleacetic acid levels in rats with experimental liver failure," *J. Pharmacobio-Dyn.*, vol. 8, no. 9, pp. 780–784, 1985.
[48] V. Bhatia et al., "Urinary potassium loss in children with acute liver failure and acute viral Hepatitis," *J. Pediatr. Gastroenterol. Nutr.*, vol. 57, no. 1, pp. 102–108, 2013.
[49] X. Li et al., "FCMDAP: Using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs," *BMC Syst. Biol.*, vol. 13, no. 2, pp. 1–16, 2019.
[50] X. Lei, J. Tie, and Y. Pan, "Inferring metabolite-disease association using graph convolutional networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 688–698, Mar./Apr. 2022.
[51] L. Liu, Y. Wei, Q. Zhang, and Q. Zhao, "SSCRB: Predicting circRNA-RBP interaction sites using a sequence and structural feature-based attention model," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1762–1772, Mar. 2024.
[52] B. Tu et al., "Hyperspectral image classification with multi-scale feature extraction," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 534.
[53] J. Zhao et al., "Predicting potential interactions between lncRNAs and proteins via combined graph auto-encoder methods," *Brief. Bioinf.*, vol. 24, no. 1, 2023, Art. no. 527.
[54] J. Wang et al., "Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints," *Methods*, vol. 221, no. 1, pp. 18–26, 2024.
[55] T. Wang, J. Sun, and Q. Zhao, "Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism," *Comput. Biol. Med.*, vol. 153, no. 1, 2023, Art. no. 64.
[56] Y. Guo et al., "Variational gated autoencoder-based feature extraction model for inferring disease-miRNA associations based on multiview features," *Neural Netw.*, vol. 165, no. 1, pp. 491–505, 2023.
[57] X. Yang et al., "Multi-task aquatic toxicity prediction model based on multi-level features fusion," *J. Adv. Res.*, vol. 1, no. 1, 2024, Art. no. 1.
[58] K. Martin et al., "The biogeographic differentiation of algal microbiomes in the upper ocean from pole to pole," *Nat. Commun.*, vol. 12, no. 1, 2021, Art. no. 5483.