frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 December 2024
DOI 10.3389/fmicb.2024.1483983

OPEN ACCESS

EDITED BY
Liang Wang,
Guangdong Provincial People's Hospital,
China

REVIEWED BY
Guanghui Li,
East China Jiaotong University, China
Lihong Peng,
Hunan University of Technology, China

*CORRESPONDENCE
Xiaogang Peng
✉ pengxg@szu.edu.cn
Xiaozheng Li
✉ li.xiaozheng@szu.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 22 August 2024
ACCEPTED 14 October 2024
PUBLISHED 13 December 2024

CITATION
Jiang C, Feng J, Shan B, Chen Q, Yang J, Wang G, Peng X and Li X (2024) Predicting microbe-disease associations via graph neural network and contrastive learning.
Front. Microbiol. 15:1483983.
doi: 10.3389/fmicb.2024.1483983

Predicting microbe-disease associations via graph neural network and contrastive learning

Cong Jiang[1,2†], Junxuan Feng[1,2†], Bingshen Shan[1,2†], Qiyue Chen[3], Jian Yang[4,5], Gang Wang[4,5], Xiaogang Peng[2]* and Xiaozheng Li[6,7]*

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China, [3]College of Management, Shenzhen University, Shenzhen, China, [4]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders and National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China, [5]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China, [6]College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China, [7]JCY Biotech Ltd., Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen, China

In the contemporary field of life sciences, researchers have gradually recognized the critical role of microbes in maintaining human health. However, traditional biological experimental methods for validating the association between microbes and diseases are both time-consuming and costly. Therefore, developing effective computational methods to predict potential associations between microbes and diseases is an important and urgent task. In this study, we propose a novel computational framework, called GCATCMDA, for forecasting potential associations between microbes and diseases. Firstly, we construct Gaussian kernel similarity networks for microbes and diseases using known microbe-disease association data. Then, we design a feature encoder that combines graph convolutional network and graph attention mechanism to learn the node features of networks, and propose a feature dual-fusion module to effectively integrate node features from each layer's output. Next, we apply the feature encoder separately to the microbe similarity network, disease similarity network, and microbe-disease association network, and enhance the consistency of features for the same nodes across different association networks through contrastive learning. Finally, we pass the microbe and disease features into an inner product decoder to obtain the association scores between them. Experimental results demonstrate that the GCATCMDA model achieves superior predictive performance compared to previous methods. Furthermore, case studies confirm that GCATCMDA is an effective tool for predicting microbe-disease associations in real situations.

KEYWORDS

microbe-disease associations, graph convolutional network, graph attention mechanism, contrastive learning, gut microbial metagenomics

1 Introduction

Microbes are primarily composed of bacteria, fungi, archaea, and viruses, predominantly inhabit the gut within the human body (Sommer and Bäckhed, 2013; Blum, 2017). The gut microbiota is closely associated with human health, playing a crucial role in regulating host physiological processes, such as immunity and metabolism (Lynch and Pedersen, 2016; Tooley, 2020). In recent years, biological experiments have demonstrated that dysbiosis or imbalance in the human microbiota could cause human diseases (Marchesi et al., 2016), such as liver diseases (Henao-Mejia et al., 2013), diabetes (Paun et al., 2017), obesity (Tseng and Wu, 2019), and even cancer (Schwabe and Jobin, 2013). However, traditional biological experiments suffer from drawbacks such as long experimental cycles and expensive costs. Therefore, if we can utilize effective computational methods to predict potential sets of associations between microbes and diseases in advance, it would be possible to reduce unnecessary experimental trials and costs in traditional biological experiments, thereby accelerating the development of research in the field of microbe-disease associations.

Current computational methods for predicting microbe-disease associations can primarily be categorized into three categories, namely network-based methods, random walk-based methods, and deep learning-based methods. The network-based methods infer the potential association between microbes and diseases by utilizing the topological information within the network. For example, Chen et al. (2017) proposed a KATZHMDA model based on the KATZ measure, which scores potential disease related microbes by calculating all paths of different lengths between microbes and diseases. Bao et al. (2017) proposed the Network Consistency Projection for Human Microbe-Disease Association Prediction (NCPHMDA) model, evaluating the association scores between microbes and diseases by computing disease space projection scores and microbe space projection scores. Long and Luo (2019) designed a meta-graph-based method named WMGHMDA, which calculates the probability scores of microbe-disease pairs by utilizing a weighted meta-graph search algorithm on a heterogeneous network. Wang et al. (2023) proposed a SAELGMDA model by combining sparse autoencoder and Light Gradient boosting machine.

The success of random walk algorithms in graph data processing has prompted researchers to propose various microbe-disease association prediction algorithms based on this approach. For instance, Zou et al. (2017) developed a novel computational model of BiRWHMDA, which predicts potential microbe-disease associations by bi-random walks on a heterogeneous network. Luo and Long (2018) proposed a novel computational model of NTSHMDA, which integrates network topology similarity into the restarted random walk algorithm to distinguish the walking probabilities of disease-microbe node pairs. Yan et al. (2019) introduced a BRWMDA method, predicting potential microbe-disease associations by executing bi-random walks with different steps on microbe and disease networks.

With the significant achievements of deep learning algorithms in various research fields, researchers have gradually begun to explore the application of these algorithms in the task of predicting the associations between microbes and diseases. For example, Ma and Jiang (2020) developed an end-to-end graph convolutional neural network-based mining model NinimHMDA to predict different types of microbe-disease associations. Long et al. (2021) proposed a novel deep learning framework of GATMDA, which utilizes graph attention networks along with inductive matrix completion for predicting human microbe-disease associations. Hua et al. (2022) developed a multi-view graph augmentation convolutional network (MVGCNMDA) to predict potential disease-associated microbes. Jiang et al. (2022) proposed the KGNMDA method, using a knowledge graph neural network method for predicting microbe-disease associations. Peng et al. (2023) developed a computational method for predicting microbe-disease associations, named GPUDMDA, which integrates graph attention autoencoder, positive-unlabeled learning, and deep neural network.

In addition to the three mainstream methods mentioned, some computational approaches for microbe-disease prediction have been developed based on regularization and matrix factorization/completion techniques. For instance, Wang et al. (2017) proposed a semi-supervised computational model of Laplacian Regularized Least Squares for Human Microbe—Disease Association (LRLSHMDA) to predict microbe-disease associations. Shen et al. (2017) developed a computational method of CMFHMDA, which utilizes collaborative matrix factorization to reconstruct correlation matrices between diseases and microbes. Liu et al. (2023) proposed a novel method called MNNMDA to predict microbe-disease associations by applying a Matrix Nuclear Norm method.

Among the methods mentioned above, network-based and random walk-based methods may encounter constraints in learning features of nodes representing microbes and diseases with few known associations, due to the limited information propagation caused by the sparsity of the microbe-disease association network. Meanwhile, matrix factorization/completion methods can only capture linear associations, thus failing to accurately capture the nonlinear interactions between microbes and diseases. Recent studies have suggested that graph neural network algorithms in deep learning could offer a more effective approach for learning node features in microbe-disease association networks. Therefore, this study further attempts to design node feature learning algorithms based on graph neural networks, aiming to obtain more effective node features from the microbe-disease association network, thereby predicting more accurate candidate sets of microbe-disease associations.

In this work, we propose a deep learning framework named GCATCMDA, which explores the application of graph neural networks for the microbe-disease association prediction task. First Gaussian kernel similarity is calculated based on known microbe-disease association data to construct microbe similarity networks and disease similarity networks. We then combine graph convolutional networks and graph attention mechanisms to learn feature representations of microbes and diseases in different networks, and propose a feature dual-fusion module to effectively integrate node features generated by each graph attention layer. Next, we utilize contrastive learning to enhance the feature consistency of the same microbe (or disease) across different association networks. Finally, the obtained microbe and disease features are inputted into an inner product decoder to compute their corresponding association scores. The model can obtain better node features through GCAT aggregation. In addition, contrastive learning increases the distance between nodes, allowing the model to better distinguish nodes and make subsequent predictions better. Experimental results demonstrate that the GCATCMDA model achieves better predictive performance compared to previous methods, and case studies of obesity and IBD (inflammatory

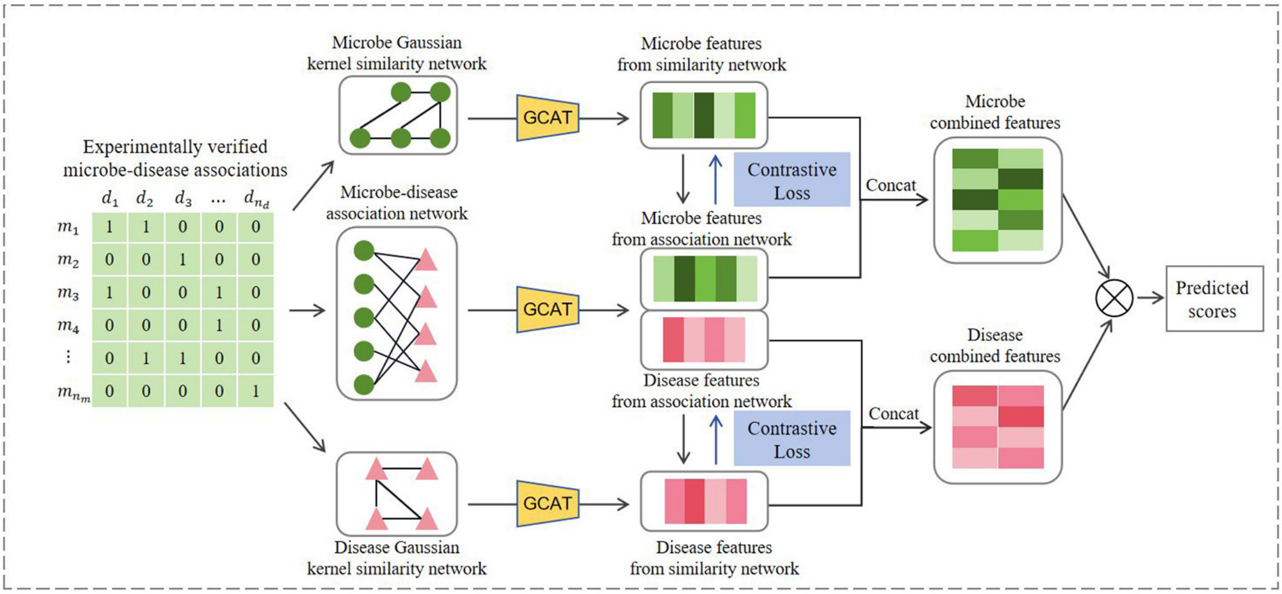

FIGURE 1
The workflow of GCATCMDA for microbe–disease prediction.

bowel disease) confirm the high accuracy of the microbe-disease association candidate set produced by our method.

2 Materials and methods

2.1 Datasets

The dataset in this study was sourced from the HMDAD database (http://www.cuilab.cn/hmdad), which collects known associations between microbes and diseases by searching past research literature (Ma et al., 2017). HMDAD adapted a systematic approach by only including associations that have been experimentally validated and published in reputable journals. This ensures a high level of reliability in the dataset. Past researchers commonly employ metagenomic sequencing techniques to analyze fluctuations in microbial community abundance within specific diseases, contrasting them with the microbial compositions of healthy individuals, thus exploring the associations between microbes and diseases. In the HMDAD dataset, a microbe-disease association pair may contain multiple entries from different research literature sources. Therefore, here, we regard the same microbe-disease association from different evidences as a pair, further removing the redundant information present in the HMDAD dataset. Finally, for this study, we employed a dataset consisting of 450 microbe-disease associations, encompassing 39 human diseases and 292 microbes.

2.2 Problem definition

For the convenience of clarity in describing the subsequent research methods, we provide a simple problem definition for the task of predicting associations between microbes and diseases here. We denote $M = \{m_1, m_2, \ldots, m_{n_m}\}$ and $D = \{d_1, d_2, \ldots, d_{n_d}\}$ as the sets representing n_m microbes and n_d diseases, respectively. The matrix $A \in \mathbb{R}^{n_m \times n_d}$ represents the known associations between microbes and diseases, where $A_{ij} = 1$ if microbe m_i is associated with disease d_j, otherwise $A_{ij} = 0$. However, $A_{ij} = 0$ does not mean that microbe m_i has no relation with disease d_j. It may be the reason that their association has not yet been discovered. Therefore, the task of predicting associations between microbes and diseases aims to find microbe m_i for each disease d_j where $A_{ij} = 0$ in the known association matrix, but microbe m_i is actually related to disease d_j.

2.3 GCATCMDA

Figure 1 illustrates the workflow of GCATCMDA, a model based on graph neural networks and contrastive learning for predicting effective candidate sets of microbe-disease associations. First microbe-microbe and disease-disease Gaussian kernel similarity networks are constructed using known associations. The model then integrates graph neural networks and contrastive learning principles to extract meaningful feature representations of microbes and diseases from the association networks. Last the obtained microbe and disease features are fed into an inner product decoder to compute their corresponding association scores. A detailed description of the key components of this model is elucidated below.

2.3.1 Microbe and disease similarity network construction

Previous study (Chen et al., 2017) have hypothesized that functionally similar microbes (or diseases) tend to exhibit similar interaction or non-interaction patterns with similar diseases (or

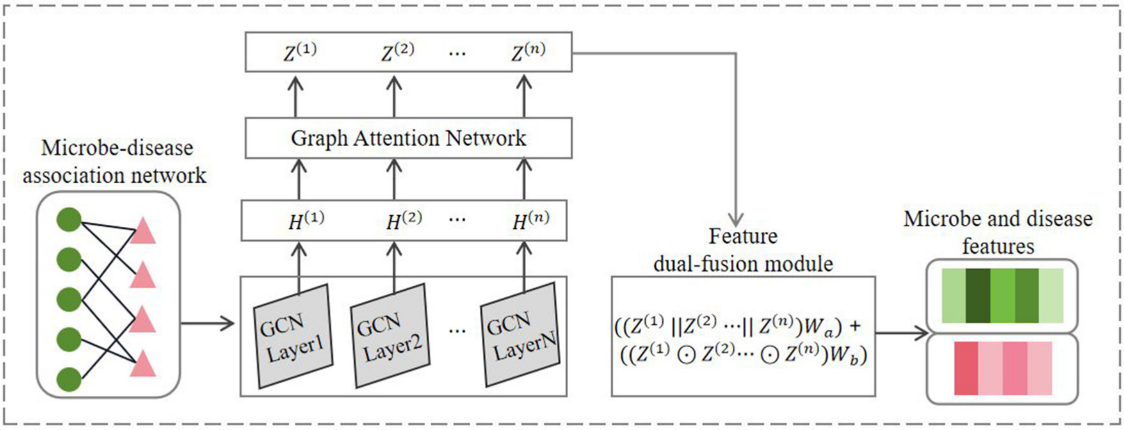
$Z^{(1)}$ $Z^{(2)}$ ··· $Z^{(n)}$
Graph Attention Network
Microbe-disease association network
$H^{(1)}$ $H^{(2)}$ ··· $H^{(n)}$
GCN Layer1
GCN Layer2
···
GCN LayerN
Feature dual-fusion module
$((Z^{(1)} \| Z^{(2)} \cdots \| Z^{(n)})W_a) + ((Z^{(1)} \odot Z^{(2)} \cdots \odot Z^{(n)})W_b)$
Microbe and disease features

FIGURE 2
The flowchart of GCAT node feature encoding in microbe-disease association network.

microbes). They utilize Gaussian kernel functions to measure the similarity between two microbes (or diseases) in the same space. Therefore, in this study, we consider constructing microbe and disease similarity networks based on Gaussian kernel similarity scores for microbes and diseases.

We have recorded the known associations between microbes and diseases using the association matrix $A \in \mathbb{R}^{n_m \times n_d}$. The calculation formulas for the Gaussian kernel similarity score between microbe m_i and m_j, and between disease d_i and d_j, are as follows:

$$\mathrm{KM}(m_i, m_j) = \exp(-\lambda_m \|\mathrm{IP}(m_i) - \mathrm{IP}(m_j)\|^2) \tag{1}$$

$$\mathrm{KD}(d_i, d_j) = \exp(-\lambda_d \|\mathrm{IP}(d_i) - \mathrm{IP}(d_j)\|^2) \tag{2}$$

where $KM(m_i, m_j)$ represents the Gaussian kernel similarity score between microbes m_i and m_j, and $KD(d_i, d_j)$ represents the Gaussian kernel similarity score between diseases d_i and d_j. The term $\mathrm{IP}(m_i)$ represents the i-th row of the association matrix A recording the associations between microbe m_i and other diseases, $\mathrm{IP}(d_i)$ represents the i-th column of the association matrix A recording the associations between disease d_i and other microbes. The parameters λ_m and λ_d represent the normalized kernel bandwidths and are defined as follows:

$$\lambda_m = \frac{\lambda'_m}{\frac{1}{n_m}\sum_{i=1}^{n_m} \|\mathrm{IP}(m_i)\|} \tag{3}$$

$$\lambda_d = \frac{\lambda'_d}{\frac{1}{n_d}\sum_{i=1}^{n_d} \|\mathrm{IP}(d_i)\|} \tag{4}$$

where n_m and n_d represented the number of microbes and diseases. And λ'_m and λ'_d are the original bandwidths, and generally both set to 1.

We consider microbes (or diseases) to be strongly associated with each other when the Gaussian kernel similarity score between microbes (or diseases) exceeds a threshold of t. Therefore, the association matrices MA for microbes and DA for diseases can be expressed as follows:

$$\mathrm{MA}(m_i, m_j) = \begin{cases} 1, & \text{if } \mathrm{KM}(m_i, m_j) \geq t \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\mathrm{DA}(m_i, m_j) = \begin{cases} 1, & \text{if } \mathrm{KD}(d_i, d_j) \geq t \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

2.3.2 GCAT

Inspired by the work of Sun et al. (2022) on predicting metabolite-disease associations, this study adopted the GCAT feature encoder. The encoder initially combines graph convolution algorithms and graph attention mechanisms to learn the nodal features of the network, followed by the design of a feature dual-fusion module to effectively integrate the node features outputted by each graph attention layer. Since the GCAT feature encoder learns embedding representations on different association networks in a similar process, we take microbe-disease association network as an example to introduce the process of learning node features, as illustrated in Figure 2.

We represent the microbe-disease association network using a symmetric adjacency matrix $G \in \mathbb{R}^{(n_m+n_d)\times(n_m+n_d)}$, where n_m and n_d denote the numbers of microbes and diseases, respectively. The initial features of nodes in the network are represented by the matrix $H^{(0)}$.

$$G = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{7}$$

$$H^{(0)} = \begin{bmatrix} KM & 0 \\ 0 & KD \end{bmatrix} \tag{8}$$

Considering the ability of graph convolutional networks in capturing the global graph structural information, and the ability of graph attention mechanisms to assign attention weights to different nodes based on the local graph structure, the GCAT feature encoder integrates these two algorithms to effectively learn the node features of the microbe-disease association network. Firstly, the GCAT feature encoder linearly projects the node features of $H^{(0)}$ onto a feature space of dimensional size F, denoted as $H^{(0)} = H^{(0)}W$, where $W \in \mathbb{R}^{(n_m+n_d)\times F}$ is the weight matrix. Next, this module employs the graph convolutional networks (GCN) proposed by Kipf and Welling (2016) to learn node features in the network. GCN learns low-dimensional representations of nodes by aggregating neighbor node information through graph convolution operations while preserving the graph's structural information. The first-layer graph convolutional propagation formula for graph G can be expressed as:

$$\mathbf{H}^{(1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{G}\tilde{D}^{-\frac{1}{2}}\mathbf{H}^{(0)}\mathbf{W}^{(0)}\right) \tag{9}$$

Here, $\sigma(.)$ denotes the activation function, $\tilde{G} = G + I$ represents the adjacency matrix with self-loops added, $\tilde{D}$ is the degree matrix of $\tilde{G}$, $\mathbf{W}^{(0)}$ denotes the trainable weight matrix of the first-layer graph convolution, and $\mathbf{H}^{(1)}$ represents the feature matrix outputted by the first-layer graph convolution.

Subsequently, the GCAT feature encoder enhances the learned node feature representations from the graph convolutional layers by incorporating a graph attention mechanism to aggregate weighted sums of neighbor information. In this study, we adopt the graph attention network (GAT) proposed by Veličković et al. (2017), which introduces an attention mechanism to assign different attention weights to the features of different neighbor nodes, enabling to focus on important neighbor features during aggregation for the target node. Thus, following the computation of the first-layer graph convolution, the attention scores $\alpha_{ij}^{(1)}$ for node j with respect to its neighbor node i in graph G can be calculated as:

$$\alpha_{ij}^{(1)} = \frac{\exp(f([W_{\text{att}}^{(1)}h_i^{(1)}||W_{\text{att}}^{(1)}h_j^{(1)}]))}{\sum_{k\in N_i}\exp(f([W_{\text{att}}^{(1)}h_i^{(1)}||W_{\text{att}}^{(1)}h_k^{(1)}]))} \tag{10}$$

where $||$ denotes the concatenation operation, $h_*^{(1)}$ represents the node features obtained by the graph G through the first-layer graph convolution, $W_{\text{att}}^{(1)}$ represents the weight matrix for the linear transformation of node features, N_i denotes the first-order neighboring nodes of node i. The attention mechanism $f(\cdot)$ is a single-layer feedforward neural network, parametrized by a weight vector $\vec{a} \in \mathbb{R}^{2F}$, and applying the LeakyReLU nonlinearity. We further employs a multi-head attention mechanism to stabilize the process of learning node representations in attention networks. It aggregates the features obtained from all attention heads by taking their average. Thus, the updated feature $z_i^{(1)}$ for node i via graph attention mechanism can be expressed as follows:

$$z_i^{(1)} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^{k}\cdot\left(W_{\text{att}}^{k}\right)^{(1)}h_j^{(1)}\right) \tag{11}$$

Here, σ denotes the activation function, K represents the number of attention heads, N_i signifies the neighborhood of node i, α_{ij}^{k} represents the attention coefficient for node j with respect to node i in the k-th attention head, W_{att}^{k} is the weight matrix for attention in the k-th head, and $h_j^{(1)}$ denotes the feature vector of node j after the first graph convolutional layer.

Finally, inspired by the work of Wang et al. (2019) on node feature fusion, this study further designs a feature dual-fusion module, which considers both concatenation and element-wise product operations to integrate the node features outputted by each graph attention layer. We posit that the concatenation operation helps preserve more node feature information, while the element-wise product operation emphasizes the correlation between node features. We demonstrated the effectiveness of this fusion module in ablation experiments. The node features outputted by each graph attention layer in the GCAT feature encoder can be represented as $\{Z^{(1)}, Z^{(2)}, \cdots, Z^{(N)}\}$. Then, the feature dual-fusion module can be represented by the following equation:

$$\begin{aligned} Z &= (Z^{(1)}||Z^{(2)}||\cdots||Z^{(N)})W_a \\ &+(Z^{(1)}\odot Z^{(2)}\odot\cdots\odot Z^{(N)})W_b \end{aligned} \tag{12}$$

Here, $||$ represents concatenation, and $\odot$ represents element-wise (Hadamard) product, $W_a \in \mathbb{R}^{(N\times F)\times F}$ and $W_b \in \mathbb{R}^{F\times F}$ denote the trainable weight matrices, Z represents the final node feature.

In summary, this study represents the final microbe and disease features obtained from the microbe-disease association network as $Z_A^m \in \mathbb{R}^{n_m\times F}$ and $Z_A^d \in \mathbb{R}^{n_d\times F}$, respectively. Similarly, the microbe features obtained from the microbe similarity network are represented as $Z_S^m \in \mathbb{R}^{n_m\times F}$, and the disease features obtained from the disease similarity network are represented as $Z_S^d \in \mathbb{R}^{n_d\times F}$.

2.3.3 Contrastive learning

Inspired by the work of Jin et al. (2024) on miRNA-disease association prediction, this study introduces contrastive learning to enhance the consistency of features of the same nodes across different association networks and the distinctiveness of features between different pairs of nodes. This approach leverages the complementary information among various association networks to obtain more effective representations of microbe and disease features. This module employs the contrastive loss function proposed by Zhu et al. (2020) for graph nodes. It considers the node features of the same disease d_i obtained from different association networks $(Z_A^{d_i}, Z_S^{d_i})$ as positive samples, while all other pairs of different nodes form negative sample pairs. Therefore, the contrastive learning loss function $Loss_d$ for disease node features across different association networks can be expressed as:

$$\begin{aligned} &l(Z_A^{d_i}, Z_S^{d_i}) = \\ &\log\left(\frac{e^{\theta(Z_A^{d_i},Z_S^{d_i})/\tau}}{e^{\theta(Z_A^{d_i},Z_S^{d_i})/\tau} + \sum_{k\neq i}(e^{\theta(Z_A^{d_i},Z_A^{d_k})/\tau} + e^{\theta(Z_A^{d_i},Z_S^{d_k})/\tau})}\right) \end{aligned} \tag{13}$$

$$\text{Loss}_d = -\frac{1}{2n_d}\sum_{i=1}^{n_d}\left[l(Z_A^{d_i}, Z_S^{d_i}) + l(Z_S^{d_i}, Z_A^{d_i})\right] \tag{14}$$

where $\theta(\cdot)$ is the cosine similarity, τ is a temperature parameter, n_d denotes the number of disease. Similarly, the contrastive learning loss function Loss_m for microbe node features across different association networks can be formulated as follows:

$$\mathrm{Loss}_m = \frac{1}{2n_m}\sum_{i=1}^{n_m}\left[l(Z_A^{m_i}, Z_S^{m_i}) + l(Z_S^{m_i}, Z_A^{m_i})\right] \tag{15}$$

where n_m denotes the number of microbe. Therefore, the overall loss function of the GCATCMDA model in the contrastive learning module is formulated as follows:

$$Loss_{\mathrm{contrast}} = (Loss_d + Loss_m) \tag{16}$$

2.3.4 Microbe—disease associations prediction

This study aggregates the node features of microbes and diseases obtained from different association networks through vector concatenation, resulting in the final microbial feature representation $Z_m = [Z_A^m || Z_S^m] \in \mathbb{R}^{n_m \times 2F}$ and disease feature representation $Z_d = [Z_A^d || Z_S^d] \in \mathbb{R}^{n_d \times 2F}$. Subsequently, these aggregated feature representations are passed into an inner product decoder to compute the association scores between microbes and diseases. The calculation process is as follows:

$$A' = \mathrm{sigmoid}(Z_m Z_d^T) \tag{17}$$

Where sigmoid is the activation function defined as $1/(1 + e^{-x})$, which maps output values to the interval (0, 1), A'_{ij} represents the association prediction score between microbe m_i and disease d_j.

Finally, the training of the GCATCMDA model employs Binary Cross-Entropy as the loss function for microbe-disease association prediction. The formula for this function is as follows:

$$\begin{aligned}\mathrm{Loss}_{\mathrm{classify}} = -\frac{1}{N}\sum_{(i,j)\in\{N^+ \cup N^-\}} & [A_{(i,j)} \log(A'_{ij}) \\ & + (1 - A_{(i,j)}) \log(1 - A'_{ij})]\end{aligned} \tag{18}$$

Where N denotes the total number of associations between microbes and diseases, N^+ represents the confirmed associations between microbes and diseases, and N^- represents the associations yet to be confirmed. The tuple (i, j) represents the association between microbe m_i and disease d_j. If (m_i, d_j) belongs to N^+, then $A_{(i,j)} = 1$; otherwise, $A_{(i,j)} = 0$. A'_{ij} denotes the predicted association score by the model for this association pair. Therefore, the overall loss function of the GCATCMDA model can be expressed as:

$$Loss_{\mathrm{total}} = Loss_{\mathrm{classify}} + \lambda Loss_{\mathrm{contrast}} \tag{19}$$

Where λ represents the weighting parameter for the contrastive learning loss. The detailed steps of GCAT to predict novel microbe—disease associations is described in Algorithm 1.

3 Results and discussion

In this section, we will provide an exposition of the experimental setup and subsequently delve into an analysis and discussion of the experimental results.

```
Input: Microbe-disease associations x, real associations y
Output: Loss value
x_micro ← Gaussion_kernel1(x)
x_disease ← Gaussion_kernel2(x)
micro_f_association, disease_f_association ← GCAT_association(x)
micro_f_similarity ← GCAT_micro_similarity(x_micro)
disease_f_similarity ← GCAT_disease_similarity(x_disease)
micro_feature ← cat([micro_f_association, micro_f_similarity], dim = −1)
disease_feature ← cat([disease_f_association, disease_f_similarity], dim = −1)
pred ← Sigmoid(micro_feature ∗ disease_feature)
loss ← Binary_Cross_Entropy(pred, y)
micro_contrastive_loss ← Contrastive_Loss(micro_f_similarity, micro_f_similarity)
disease_contrastive_loss ← Contrastive_Loss(disease_f_similarity, disease_f_association)
loss ← loss + micro_contrastive_loss + disease_contrastive_loss
return loss
```

Algorithm 1. GCAT framework for microbe-disease association.

3.1 Experimental setup

The GCATCMDA model proposed in this study is a microbe-disease association prediction model based on graph neural networks and contrastive learning. It aims to predict potential associations between microbes and diseases from known microbial-disease association dataset. The hyperparameter settings required for this model are described as follows. Firstly, the Gaussian kernel similarity threshold t needs to be set for constructing microbe and disease similarity networks. Secondly, parameters need to be set for the GCAT feature encoder module, including the dimensionality F of node features, the number of network layers L for graph convolution, and the number of attention heads *heads* for the graph attention mechanism. Then, in the contrastive learning loss module, the temperature hyperparameter τ and the weight parameter λ relative to the total loss are adjusted. Finally, the GCATCMDA model is trained using the Adam (Kingma and Ba, 2014) optimizer, with parameters including the learning rate *lr*, weight decay *wd*, and the number of training iterations *epochs*.

This study determines the optimal parameter settings of the GCATCMDA model on the dataset by enumerating different parameter combinations. Subsequently, there is an analysis of key parameters t, F, L, and *heads*. After comparing experimental results, the optimal hyperparameter settings for the GCATCMDA model on the HMDAD dataset are determined as follows: $t = 0.4$, $F = 128$, $L = 3$, $heads = 2$, $\tau = 1$, $\lambda = 0.2$, $lr = 0.00001$, $wd = 0.001$, and $epochs = 100$.

In order to verify the effectiveness of the proposed GCATCMDA model, we compares it with KATZHMDA (Chen et al., 2017), LRLSHMDA (Wang et al., 2017), NTSHMDA (Luo and Long, 2018), and KGNMDA (Jiang et al., 2022). These five methods are recognized for their outstanding performance in this task in past studies and provide research methods with

TABLE 1 Classification performance comparison of GCATCMDA with existing methods.

Cross-validation	Methods	AUC	AUPR
Five-fold-CV	KATZHMDA	0.877 (0.023)	0.890 (0.021)
	LRLSHMDA	0.801 (0.032)	0.774 (0.039)
	NTSHMDA	0.892 (0.028)	0.892 (0.036)
	KGNMDA	0.895 (0.021)	0.903 (0.020)
	GCATCMDA	**0.908 (0.020)**	**0.913 (0.022)**
10-fold-CV	KATZHMDA	0.880 (0.031)	0.892 (0.027)
	LRLSHMDA	0.805 (0.047)	0.788 (0.058)
	NTSHMDA	0.897 (0.030)	0.897 (0.038)
	KGNMDA	0.900 (0.029)	0.909 (0.029)
	GCATCMDA	**0.910 (0.026)**	**0.914 (0.033)**

The values in bold represent the best ones.

open-source code. For negative samples required in model training, this study randomly selects an equal number of negative samples from all unknown microbe-disease association pairs. The number of negative samples matched the number of positive samples as to maintain a balanced dataset. In each cross-validation experiment, the Gaussian kernel similarity scores for microbes and diseases are recalculated based on the training set to ensure the effectiveness of evaluating model performance through the test set. In this experiment, we employ the same dataset and follow the hyperparameter settings used in the original papers or provide open source codes for other compared models. We adopted the same evaluation metrics as the previous study (Jiang et al., 2022), including the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPR) to assess the performance of the models. To evaluate the performance of these models in predicting potential associations between microbes and diseases, this study conducted 10 repetitions of five-fold cross-validation experiments and 10 repetitions of ten-fold cross-validation experiments by setting different random seeds, and then computed the average to ensure the accuracy of our results.

3.2 The classification performance of models

The comparative results of the two cross-validation experiments conducted on the HMDAD dataset for the five aforementioned models are presented in Table 1. The optimal performance is highlighted in bold, with standard deviations indicated in parentheses. To provide readers with a clearer visualization of the performance of the models, this study further plotted the ROC curve and PR curve, as shown in Figures 3, 4, respectively.

From the experimental data presented above, it can be observed that the GCATCMDA model proposed in this study has achieved excellent predictive performance in the task of predicting associations between microbes and diseases, surpassing methods proposed in previous studies. For instance, in the five-fold cross-validation experiment, the model obtained an approximate 1.3% improvement in AUC compared to the best previous predictive performance. Similarly, in the ten-fold cross-validation experiment, the model obtained an approximate 1.0% enhancement in AUC compared to the best previous predictive performance. The improvement in predictive performance was slightly more pronounced in the fold-fold cross-validation compared to the 10-fold cross-validation. This can be attributed to the larger training sets used in the 10-fold validation, which reduce variability across folds and provide more comprehensive data for model training. However, the reduced variability can lead to subtler improvements in performance metrics, as the model benefits from a more stable but less varied dataset. In contrast, the five-fold validation, with its larger test sets, introduces more variability, making performance improvements more apparent.

Graph transformer models offer strong capabilities in capturing global node features through their self-attention mechanisms (Li et al., 2024a,b). This allows them to handle complex and non-local structures, which can be beneficial for highly heterogeneous datasets. However, these models come with significant computational complexity, scaling quadratically with the number of nodes, making them less practical for large datasets like microbe-disease networks.

While the GCATCMDA model combines GCN and GAT to effectively capture both local features and selective attention on relevant neighbors, Graph Transformer models are designed to capture these relationships on a broader scale. The full attention mechanism of Graph Transformers allows them to dynamically weigh the importance of distant nodes, offering more flexibility in feature extraction across large and complex networks. In contrast, our GCATCMDA model, which combines GCNs and GATs, is more computationally efficient and particularly suited to smaller, sparser datasets like the HMDAD database. While graph transformers excel in capturing global relationships, our approach balances local feature aggregation and attention, offering a more efficient solution. Future work could explore integrating graph transformers to leverage their global feature-capturing capabilities alongside our model's efficiency in handling localized data.

3.3 Parameter analysis

The GCATCMDA model proposed in this study possesses several crucial parameters, such as the Gaussian kernel similarity threshold t for constructing microbe and disease similarity networks, the dimensionality F of node features, the number of network layers L for graph convolution, and the number of attention heads *heads* for the graph attention mechanism. Therefore, this study conducted training with different parameter combinations on the HMDAD dataset and utilized the experimental results from 10 repetitions of five-fold cross-validation to analyze the impact of these parameters on the model's performance.

As shown in Figure 5, the model fails to achieve the best predictive performance when the Gaussian kernel similarity threshold t is either set too high or too low, the optimal predictive performance of the model is attained when $t = 0.4$. Moreover,

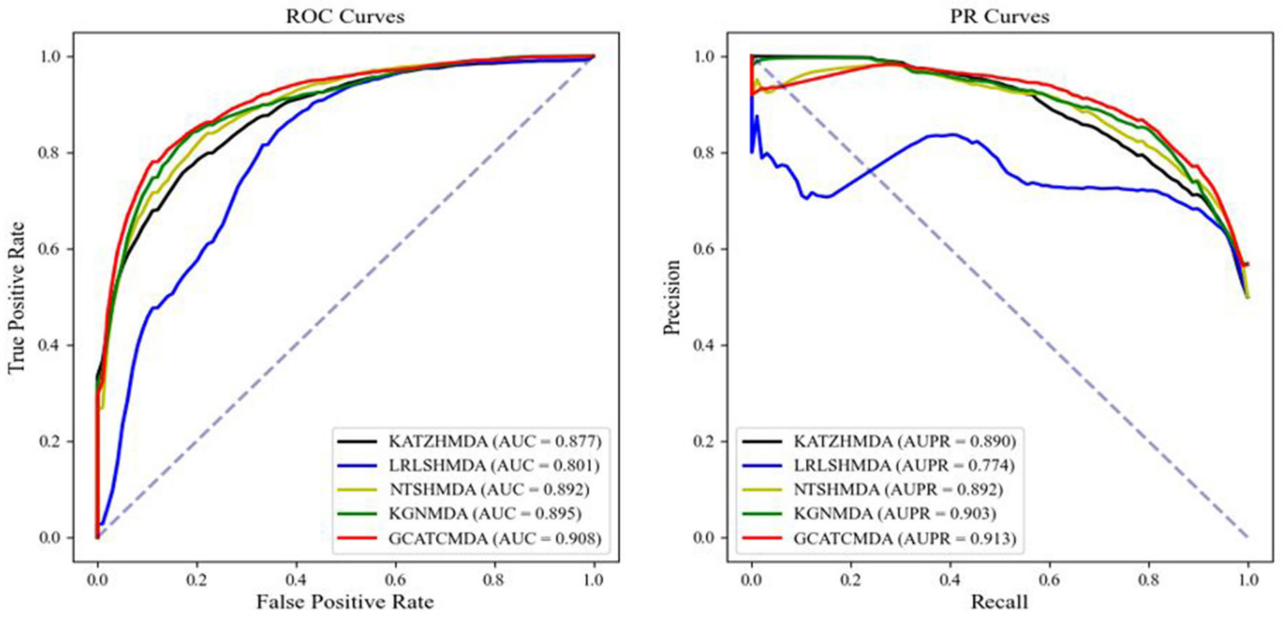

FIGURE 3
The AUROC curve and AUPR curve of five-fold CV on the HMDAD datasets between different methods.

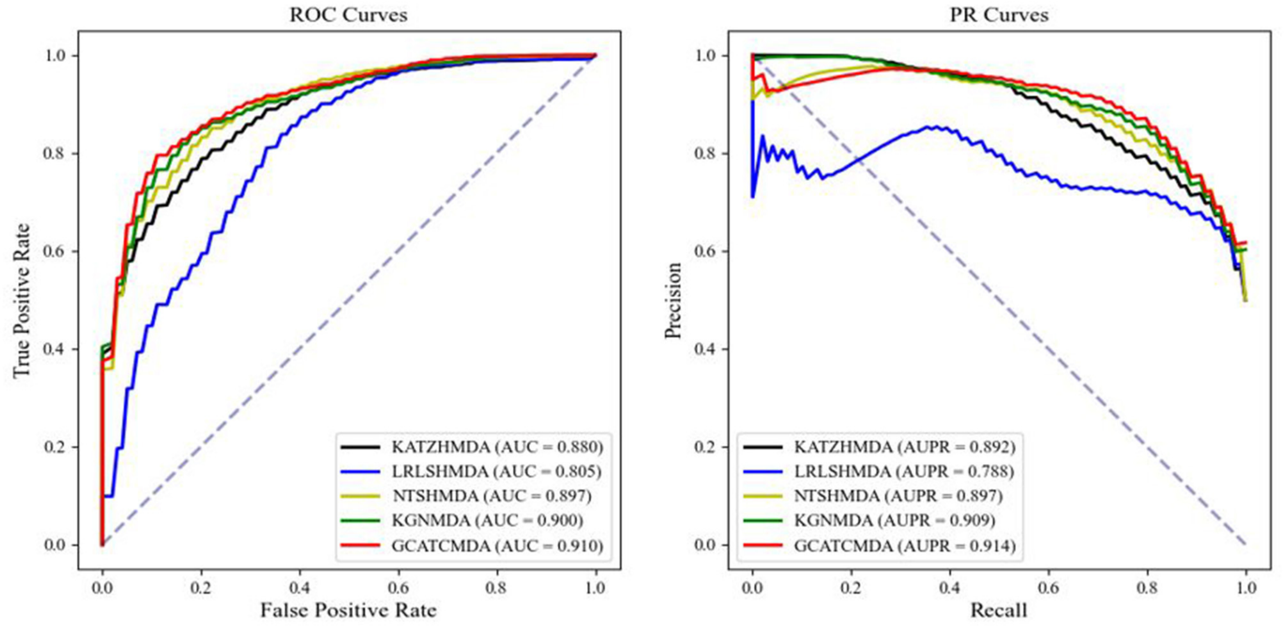

FIGURE 4
The AUROC curve and AUPR curve of 10-fold CV on the HMDAD datasets between different methods.

as the dimensionality of node features increases, the predictive performance of the model gradually improves, with the best performance observed when $F = 128$. Additionally, the model exhibits its best predictive performance when the number of network layers for graph convolution $L = 3$. Furthermore, it is observed that the evaluation metrics AUC and AUPR attain their maximum values when the number of attention heads for the graph attention mechanism $heads = 2$.

3.4 Ablation studies

To further validate the impact of each module in the GCATCMDA model on the prediction performance of microbe-disease associations, this study conducted ablation experiments on the HMDAD dataset. The evaluation metrics included AUC, AUPR, Precision, Recall, and F1 score. These metrics aimed to comprehensively analyze the influence of different modules on

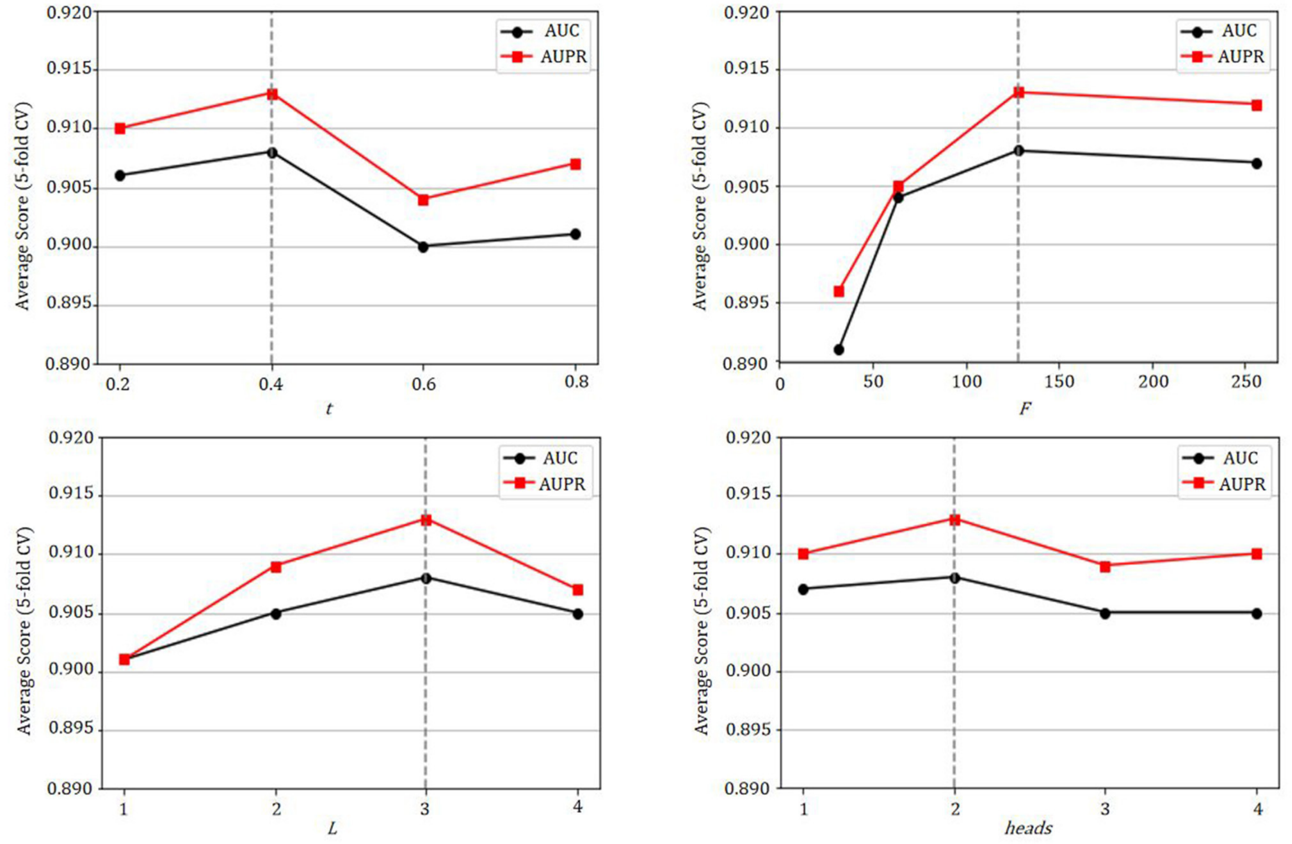

FIGURE 5
The effect of parameters *t*, *F*, *L*, and *heads* on the GCATCMDA model.

the performance of the GCATCMDA model. The experimental results represent the average scores of 10 repetitions of five-fold cross-validation experiments. Initially, given that the GCATCMDA model simultaneously utilizes microbe similarity networks, disease similarity networks, and microbe-disease association networks to learn the feature representations of microbes and diseases, this study assessed the impact of node features from different association networks on the model's prediction performance. The experimental results are illustrated in Figure 6, where GCATCMDA_sim denotes learning the feature representations of microbes and diseases only from microbe and disease similarity networks, while GCATCMDA_asso denotes learning the feature representations only from the microbe-disease association network. It can be observed from Figure 6 that integrating feature representations of microbes and diseases from different association networks effectively enhances the model's predictive performance.

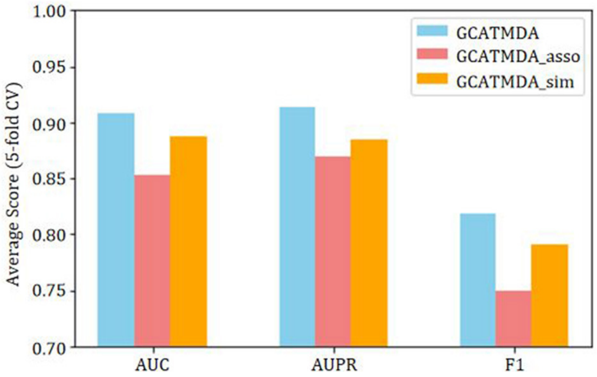

FIGURE 6
Effect of node embedding extracted from different networks on prediction.

Next, given that the GCATCMDA model mainly consists of GCN, GAT, feature dual fusion module, and contrastive learning module, this study attempted to remove each module individually to investigate the impact of different modules on the model's prediction performance. The experimental results are presented in Table 2, where "GCATCMDA_GCN" denotes the removal of the graph convolutional network from the original model, "GCATCMDA_GAT" denotes the removal of the graph attention mechanism, "GCATCMDA_SUM" denotes replacing the feature dual fusion module of the original model with a simple summation operation, and "GCATCMDA_CL" denotes the removal of the contrastive learning module from the original model. From the results in Table 2, it can be observed that both "GCATCMDA_GCN" and "GCATCMDA_GAT" exhibit lower predictive performance compared to the original model, indicating

TABLE 2 Classification performance comparison of GCATCMDA with existing methods.

Method	AUC	AUPR	Precision	Recall	F1
GCATCMDA_GCN	0.893 (0.023)	0.903 (0.021)	0.858 (0.033)	0.771 (0.041)	0.812 (0.027)
GCATCMDA_GAT	0.884 (0.034)	0.900 (0.033)	0.865 (0.040)	0.737 (0.086)	0.793 (0.062)
GCATCMDA_SUM	0.894 (0.028)	0.887 (0.040)	0.866 (0.045)	0.749 (0.078)	0.801 (0.056)
GCATCMDA_CL	0.904 (0.021)	0.908 (0.025)	0.869 (0.034)	0.770 (0.039)	0.816 (0.028)
GCATCMDA	**0.908 (0.020)**	**0.913 (0.022)**	**0.874 (0.034)**	**0.772 (0.045)**	**0.819 (0.032)**

The values in bold represent the best ones.

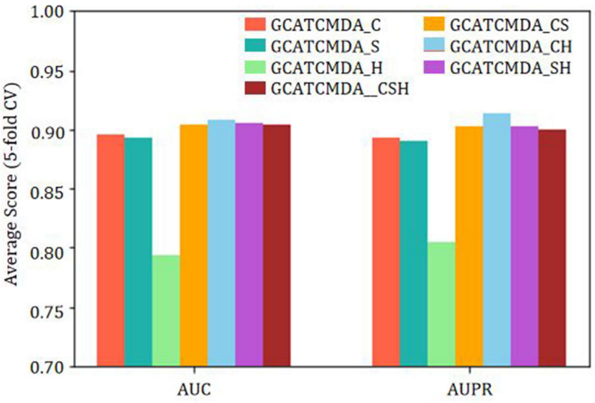

FIGURE 7
The impact of different feature fusion methods on model prediction performance.

that the integration of graph convolutional networks and graph attention mechanisms for node feature learning is effective in obtaining more informative node feature representations from the network. The predictive performance of "GCATCMDA_SUM" is also lower than that of the original model, suggesting that the designed feature dual fusion module effectively fuses node feature information outputted by the graph attention layers. Similarly, the predictive performance of "GCATCMDA_CL" is slightly lower than that of the original model, indicating that the addition of the contrastive learning module can improve the model's predictive performance to some extent.

Finally, to investigate the impact of different operations for fusing node features outputted by the graph attention layers on the GCATCMDA model prediction performance, this study sophisticatedly combined three common feature vector operations: concatenation, sum, and element-wise product. The combined fusion feature formulas are similar to the feature dual fusion formula described earlier. The experimental results are illustrated in Figure 7. GCATCMDA_C represents the use of concatenation only, GCATCMDA_S represents the use of sum only, and GCATCMDA_H represents the use of element-wise product only. GCATCMDA_CS represents the combination of concatenation and sum, GCATCMDA_CH represents the combination of concatenation and element-wise product, GCATCMDA_SH represents the combination of sum and element-wise product, and GCATCMDA_CSH represents the combination of concatenation, sum, and element-wise product. From the experimental results in Figure 7, it can be observed that selecting the combination operations of concatenation and element-wise product in the feature dual fusion module can most effectively fuse node features outputted by the graph attention layers.

3.5 Case studies

To further validate whether the GCATCMDA model can predict associations between microbes and diseases, this study initially trained the model using all known microbial-disease associations in the HMDAD dataset. Subsequently, obesity and inflammatory bowel disease (IBD), two common diseases, were selected as subjects for case analysis. The model predicted microbial associations with obesity and IBD by sorting the predicted association scores from high to low and retaining the top 20 unknown microbial associations with high scores for these two diseases. Finally, employing a literature search approach, this study validated whether these microbial associations with diseases existed by examining relevant publications in the biomedical literature database PubMed. This validation process aimed to assess the accuracy of the microbial-disease associations predicted by the GCATCMDA model.

From Table 3, it can be observed that among the top 20 associated microbes identified by the GCATCMDA model for obesity, 16 of them have been previously documented in the literature to be associated with obesity. For instance, Xu et al. (2022), by reviewing literature on gut microbiota and obesity, identified an association between Prevotella and obesity. Baradaran et al. (2021) experimentally demonstrated that individuals positive for *Helicobacter pylori* infection are more likely to suffer from obesity, with an increased risk of *Helicobacter pylori* infection among obese individuals. From Table 4, it can be observed that in IBD, among the top 20 associated microbes identified by the GCATCMDA model, 15 have been previously demonstrated to be associated with IBD in the literature. For example, Quaglio et al. (2022) demonstrated that the abundance of Bacteroidetes and Firmicutes in patients with IBD undergoes significant changes. Cardoneanu et al. (2021) experimental research showed a significant decrease in the abundance of Clostridium coccoides in patients with IBD compared to healthy individuals.

In summary, it can be observed from Tables 3, 4 that the GCATCMDA model achieves an accuracy of over 75% in predicting potential associated microbes for both obesity and inflammatory bowel disease. Therefore, this study concludes that the GCATCMDA model can provide effective and accurate candidate sets of microbes associated with diseases,

TABLE 3 Candidate microbes related to obesity predicted by GCATCMDA model.

Rank	Microbe	Evidence	Rank	Microbe	Evidence
1	*Prevotella*	PMID:35093025	11	*Enterobacter aerogenes*	Unconfirmed
2	Proteobacteria	PMID:31197613	12	*Enterobacter hormaechei*	Unconfirmed
3	*Helicobacter pylori*	PMID:34243821	13	*Klebsiella pneumoniae*	PMID:31921729
4	Lachnospiraceae	PMID:31397240	14	*Shigella dysenteriae*	Unconfirmed
5	Actinobacteria	PMID:19043404	15	*Haemophilus*	PMID:31976177
6	*Staphylococcus*	PMID:29667480	16	*Clostridium coccoides*	PMID:29667480
7	*Enterococcus*	PMID:35967777	17	*Betaproteobacteria*	Unconfirmed
8	*Clostridium*	PMID:29667480	18	*Clostridium leptum*	PMID:36756620
9	*Clostridium difficile*	PMID:25638400	19	Bacteroidales	PMID:33407104
10	*Faecalibacterium prausnitzii*	PMID:23985870	20	*Enterococcus faecium*	PMID:36590404

TABLE 4 Candidate microbes related to IBD predicted by GCATCMDA model.

Rank	Microbe	Evidence	Rank	Microbe	Evidence
1	Bacteroidetes	PMID:36157114	11	*Enterobacter hormaechei*	Unconfirmed
2	Firmicutes	PMID:36157114	12	*Klebsiella pneumoniae*	PMID:36436756
3	*Clostridium coccoides*	PMID:33548121	13	*Shigella dysenteriae*	Unconfirmed
4	*Helicobacter pylori*	PMID:30237392	14	*Clostridium leptum*	PMID:33548121
5	*Prevotella*	PMID:38053528	15	Lysobacter	Unconfirmed
6	*Clostridium difficile*	PMID:31698044	16	Rickettsiales	Unconfirmed
7	*Staphylococcus*	PMID:31662859	17	*Streptococcus mitis*	PMID:30796823
8	*Staphylococcus aureus*	PMID:31698044	18	*Xanthomonas*	PMID:35689701
9	*Enterococcus*	PMID:32292819	19	Enterobacteriaceae	PMID:24629344
10	*Enterobacter aerogenes*	Unconfirmed	20	*Lactobacillus*	PMID:37773196

thereby reducing the research costs and duration of traditional biological experiments.

4 Conclusion

This article primarily introduces the GCATCMDA model proposed in this study, aimed at predicting potential sets of microbe-disease associations based on known microbe-disease association data. Initially, the article outlines the construction of Gaussian kernel similarity networks for microbes and diseases using known association data and explains how the model combines graph neural networks with contrastive learning to obtain effective feature representations for microbes and diseases. Subsequently, experimental evaluations are conducted to compare the GCATCMDA model with existing methods, demonstrating its superiority in microbe-disease association prediction tasks. Additionally, parameter analysis experiments validate the rationality of parameter settings in the GCATCMDA model, while ablation experiments confirm the effectiveness of each module in the model. Finally, obesity and inflammatory bowel disease are selected as case studies to validate the high accuracy of the microbe-disease association candidate sets predicted by the GCATCMDA model.

The proposed model combines GCN and GAT to leverage the strengths of both approaches. GCN effectively captures local neighborhood information by performing convolution operations over graph structures, allowing the model to aggregate features across connected nodes. However, GCN applies equal weighting to all neighboring nodes, which may limit its ability to differentiate between more and less important neighbors. To address this limitation, GAT introduces an attention mechanism that assigns different importance to neighboring nodes by computing attention coefficients. This allows the model to focus more on the relevant nodes, improving its ability to capture complex interactions. By combining GCN's ability to aggregate global structural information with GAT's selective attention on important neighbors, the proposed model effectively captures both local and global patterns within the graph, leading to enhanced predictive performance.

While our study has demonstrated the effectiveness of the GCATCMDA model in predicting microbe-disease associations, there are several limitations that must be acknowledged. First, the model has only been evaluated using the HMDAD database, and its generalization ability requires further validation across other public datasets, such as HMDA and Disbiome. The limited volume of data in this study may also hinder the model's ability to capture complex patterns, suggesting the need for more extensive datasets to enhance its predictive performance. Additionally, our current approach does not differentiate between positive and negative association information, a distinction that will be addressed in future research to refine prediction accuracy. By overcoming these limitations, we anticipate further improvements in the model's robustness and its potential application across a broader range of microbial and disease studies.

In conclusion, this study asserts that the GCATCMDA model can advance the development of deep learning algorithms in the field of microbe-disease association prediction. Moreover, it effectively aids biologists in exploring potential associations between microbes and human diseases from a big data perspective, thereby reducing the costs of traditional biological experiments and accelerating research progress in the field of gut microbes and disease association studies.

Data availability statement

The HMDAD database is available at: http://www.cuilab.cn/hmdad. The source code is available upon reasonable request to the corresponding authors.

Author contributions

CJ: Investigation, Software, Visualization, Writing – original draft. JF: Formal analysis, Validation, Writing – original draft. BS: Formal analysis, Validation, Writing – original draft. QC: Data curation, Writing – original draft. JY: Funding acquisition, Writing – review & editing. GW: Resources, Writing – review & editing. XP: Methodology, Supervision, Writing – review & editing. XL: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation Project of China (82171526), Beijing Talents Project (2020A38), and Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-19).

Acknowledgments

The authors thank Shenzhen University Big Data Research Center for computing hardware facilities.

Conflict of interest

XL was a co-founder of JCY Biotech Ltd.

The remaining 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

Bao, W., Jiang, Z., and Huang, D.-S. (2017). Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinformatics* 18, 173–181. doi: 10.1186/s12859-017-1968-2

Baradaran, A., Dehghanbanadaki, H., Naderpour, S., Pirkashani, L. M., Rajabi, A., Rashti, R., et al. (2021). The association between *Helicobacter pylori* and obesity: a systematic review and meta-analysis of case-control studies. *Clin. Diabetes Endocrinol.* 7, 1–11. doi: 10.1186/s40842-021-00131-w

Blum, H. E. (2017). The human microbiome. *Adv. Med. Sci.* 62, 414–420. doi: 10.1016/j.advms.2017.04.005

Cardoneanu, A., Mihai, C., Rezus, E., Burlui, A., Popa, I., Prelipcean, C. C., et al. (2021). Gut microbiota changes in inflammatory bowel diseases and ankylosing spondylitis. *J. Gastrointestin. Liver Dis.* 30, 46–54. doi: 10.15403/jgld-2823

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G.-Y., and Wang, X.-S. (2017). A novel approach based on Katz measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

Henao-Mejia, J., Elinav, E., Thaiss, C. A., Licona-Limon, P., and Flavell, R. A. (2013). Role of the intestinal microbiome in liver disease. *J. Autoimmun.* 46, 66–73. doi: 10.1016/j.jaut.2013.07.001

Hua, M., Yu, S., Liu, T., Yang, X., and Wang, H. (2022). Mvgcnmda: multi-view graph augmentation convolutional network for uncovering disease-related microbes. *Interdiscip. Sci. Comput. Life sci.* 14, 669–682. doi: 10.1007/s12539-022-00514-2

Jiang, C., Tang, M., Jin, S., Huang, W., and Liu, X. (2022). Kgnmda: a knowledge graph neural network method for predicting microbe-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 20, 1147–1155. doi: 10.1109/TCBB.2022.3184362

Jin, Z., Wang, M., Tang, C., Zheng, X., Zhang, W., Sha, X., et al. (2024). Predicting mirna-disease association via graph attention learning and multiplex adaptive modality fusion. *Comput. Biol. Med.* 169:107904. doi: 10.1016/j.compbiomed.2023.107904

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* [Preprint]. arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv* [Preprint]. arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907

Li, G., Bai, P., Chen, J., and Liang, C. (2024a). Identifying virulence factors using graph transformer autoencoder with esmfold-predicted structures. *Comput. Biol. Med.* 170:108062. doi: 10.1016/j.compbiomed.2024.108062

Li, G., Bai, P., Liang, C., and Luo, J. (2024b). Node-adaptive graph transformer with structural encoding for accurate and robust lncrna-disease association prediction. *BMC Genomics* 25:73. doi: 10.1186/s12864-024-09998-2

Liu, H., Bing, P., Zhang, M., Tian, G., Ma, J., Li, H., et al. (2023). Mnnmda: predicting human microbe-disease association via a method to minimize matrix nuclear norm. *Comput. Struct. Biotechnol. J.* 21, 1414–1423. doi: 10.1016/j.csbj.2022.12.053

Long, Y., and Luo, J. (2019). Wmghmda: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. *BMC Bioinformatics* 20, 1–18. doi: 10.1186/s12859-019-3066-0

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe-disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform*. 22:bbaa146. doi: 10.1093/bib/bbaa146

Luo, J., and Long, Y. (2018). Ntshmda: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 17, 1341–1351. doi: 10.1109/TCBB.2018.2883041

Lynch, S. V., and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *N. Engl. J. Med*. 375, 2369–2379. doi: 10.1056/NEJMra1600266

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinform*. 18, 85–97. doi: 10.1093/bib/bbw005

Ma, Y., and Jiang, H. (2020). Ninimhmda: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe-disease association. *Bioinformatics* 36, 5665–5671. doi: 10.1093/bioinformatics/btaa1080

Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D., Hirschfield, G. M., Hold, G., et al. (2016). The gut microbiota and host health: a new clinical frontier. *Gut* 65, 330–339. doi: 10.1136/gutjnl-2015-309990

Paun, A., Yau, C., and Danska, J. S. (2017). The influence of the microbiome on type 1 diabetes. *J. Immunol*. 198, 590–595. doi: 10.4049/jimmunol.1601519

Peng, L., Huang, L., Tian, G., Wu, Y., Li, G., Li, Z., et al. (2023). Predicting potential microbe-disease associations with graph attention autoencoder, positive-unlabeled learning, and deep neural network. *Front. Microbiol*. 14:1244527. doi: 10.3389/fmicb.2023.1244527

Quaglio, A. E. V., Grillo, T. G., De Oliveira, E. C. S., Di Stasi, L. C., and Sassaki, L. Y. (2022). Gut microbiota, inflammatory bowel disease and colorectal cancer. *World J. Gastroenterol*. 28, 4053. doi: 10.3748/wjg.v28.i30.4053

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

Shen, Z., Jiang, Z., and Bao, W. (2017). "CMFHMDA: collaborative matrix factorization for human microbe-disease association prediction," in *Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part II 13* (Cham: Springer), 261–269. doi: 10.1007/978-3-319-63312-1_24

Sommer, F., and Bäckhed, F. (2013). The gut microbiota–masters of host development and physiology. *Nat. Rev. Microbiol*. 11, 227–238. doi: 10.1038/nrmicro2974

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief. Bioinform*. 23:bbac266. doi: 10.1093/bib/bbac266

Tooley, K. L. (2020). Effects of the human gut microbiota on cognitive performance, brain structure and function: a narrative review. *Nutrients* 12:3009. doi: 10.3390/nu12103009

Tseng, C.-H., and Wu, C.-Y. (2019). The gut microbiome in obesity. *J. Formosan Med. Assoc*. 118, S3–S9. doi: 10.1016/j.jfma.2018.07.009

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. (2017). Graph attention networks. *arXiv* [Preprint]. arXiv:1710.10903. doi: 10.48550/arXiv.1710.10903

Wang, F., Huang, Z.-A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction. *Sci. Rep*. 7:7601. doi: 10.1038/s41598-017-08127-2

Wang, F., Yang, H., Wu, Y., Peng, L., and Li, X. (2023). Saelgmda: identifying human microbe-disease associations based on sparse autoencoder and lightgbm. *Front. Microbiol*. 14:1207209. doi: 10.3389/fmicb.2023.1207209

Wang, X., He, X., Cao, Y., Liu, M., and Chua, T.-S. (2019). "KGAT: knowledge graph attention network for recommendation," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery* & *data mining* (New York, NY: ACM), 950–958. doi: 10.1145/3292500.3330989

Xu, Z., Jiang, W., Huang, W., Lin, Y., Chan, F. K., Ng, S. C., et al. (2022). Gut microbiota in patients with obesity and metabolic disorders–a systematic review. *Genes Nutr*. 17:2. doi: 10.1186/s12263-021-00703-6

Yan, C., Duan, G., Wu, F.-X., Pan, Y., and Wang, J. (2019). Brwmda: predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 17, 1595–1604. doi: 10.1109/TCBB.2019.2907626

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., et al. (2020). Deep graph contrastive representation learning. *arXiv* [Preprint]. arXiv:2006.04131. doi: 10.48550/arXiv.2006.04131

Zou, S., Zhang, J., and Zhang, Z. (2017). A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PLoS ONE* 12:e0184394. doi: 10.1371/journal.pone.0184394