Dual graph-embedded fusion network for predicting potential microbe-disease associations with sequence learning

Junlong Wu[1], Liqi Xiao[1], Liu Fan[1], Lei Wang[2]* and Xianyou Zhu[1,3]*

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, [2]Technology Innovation Center of Changsha, Changsha University, Changsha, China, [3]Hunan Engineering Research Center of Cyberspace Security Technology and Applications, Hengyang Normal University, Hengyang, China

Recent studies indicate that microorganisms are crucial for maintaining human health. Dysbiosis, or an imbalance in these microbial communities, is strongly linked to a variety of human diseases. Therefore, understanding the impact of microbes on disease is essential. The DuGEL model leverages the strengths of graph convolutional neural network (GCN) and graph attention network (GAT), ensuring that both local and global relationships within the microbe-disease association network are captured. The integration of the Long Short-Term Memory Network (LSTM) further enhances the model's ability to understand sequential dependencies in the feature representations. This comprehensive approach allows DuGEL to achieve a high level of accuracy in predicting potential microbe-disease associations, making it a valuable tool for biomedical research and the discovery of new therapeutic targets. By combining advanced graph-based and sequence-based learning techniques, DuGEL addresses the limitations of existing methods and provides a robust framework for the prediction of microbe-disease associations. To evaluate the performance of DuGEL, we conducted comprehensive comparative experiments and case studies based on two databases, HMDAD, and Disbiome to demonstrate that DuGEL can effectively predict potential microbe-disease associations.

KEYWORDS

long and short-term memory networks, graph attention networks, microbe-disease associations, graph convolutional neural networks, full connectivity

1 Introduction

Microorganisms play an important and far-reaching role in human life and greatly impact our health (Liang et al., 2018). Recent reports indicate that the human body is host to trillions of microorganisms (Hoffmann et al., 2016) and that the number of microorganisms in the human body far exceeds the number of human cells (Bocci, 1992). These microorganisms constitute the microbiota in the human body (Zhu et al., 2010). The microbiome plays a critical role in human physiology (Heintz-Buschart and Wilmes, 2018), helping the body's intestinal tract to reduce the growth of pathogenic bacteria and infections and synthesizing some of the vitamins and amino acids needed by the body (Islam et al., 2022). Suppose the microbial community in the human body is out of balance. In that case, it can impair the function of the immune system, increase the risk of infection with pathogens (Pickard et al., 2017), lead to malnutrition or nutritional deficiencies (Burr et al.,

2020), and contribute to the development of mental health-related problems such as anxiety and depression (Anisman et al., 2018), as well as metabolic diseases such as obesity and diabetes (Sanz et al., 2015). Of course, the microbiota can help the body regulate and prevent attacks from bacteria outside the body (Barr, 2017); for example, actinomycetes are a class of antibiotic-producing bacteria that produce a wide range of antibiotics such as streptomycin and tetracycline (Grasso et al., 2016). These antibiotics inhibit the growth of other pathogenic microorganisms and help protect the body from infection (Jagannathan et al., 2021). Therefore, predicting potential associations between microorganisms and diseases is vital for unraveling the complex mechanisms of disease occurrence and discovering potential biomarkers (Montaner et al., 2020). By inferring the interactions between microorganisms and diseases, we can better understand the diagnosis and prognosis of diseases and provide new ideas and methods for preventing, diagnosing, and treating diseases (Malla et al., 2019). As technology advances, we no longer rely solely on traditional biological methods to explore the association between microbes and disease (Gilbert et al., 2016). Instead, we are increasingly introducing computational modelling into our research to predict the role of microbes in disease occurrence, development, and treatment through techniques such as big data analytics and deep learning (Marcos-Zambrano et al., 2021), which are more practical and accurate (Najafabadi et al., 2015). Researchers have recently established a series of microbe-disease association databases to conduct an in-depth study of the potential link between microbes and diseases (Jin et al., 2022). These databases combine a large amount of microbial composition data and disease information. For example, the HMDAD database created by Ma et al. became the first to document human microbe-disease associations by manually organizing a large amount of public literature (Ma et al., 2017). This database covers 483 pieces of information about the association between 39 diseases and 292 microorganisms. Second, Janssens et al. created a microbial-disease association database called Disbiome by collecting 10,922 experimental records from 1,191 documents containing 372 diseases and 1,622 microorganisms (Janssens et al., 2018).

Researchers can explore and discover the relationships between microorganisms and different diseases using the above microbe-disease association database as the primary data. Moreover, these recent technological tools can be broadly categorized into four types, namely, network-based methods (Wu et al., 2018), matrix decomposition-based methods (Shen et al., 2021), traditional machine learning-based methods (Afshari et al., 2022), and graph neural network-based methods (Li et al., 2023).

In DuGEL, we use both Graph Convolutional Neural Network (GCN) and Graph Attention Network (GAT), where GCN is specifically designed to process graph data (Jin et al., 2021). GCN can learn feature representations at both node and graph levels, and it achieves the task of learning and predicting the representations of graph data by efficiently exploiting the connectivity relationships between the nodes (Zhou et al., 2023). By adaptively learning the attention weights between each node and its neighbouring nodes, GAT can better capture local structural information in graph data. The GAT introduction enriches the representational capabilities of the graph neural network, allowing it to perform well when dealing with

TABLE 1 The statistics of the two databases.

Datasets	Microbes	Diseases	Associations
HMDAD	292	39	450
Disbiome	1,052	218	4,351

complex graph-structured data (Munikoti et al., 2023). DuGEL can adapt to an extensive range of datasets with solid robustness.

Unlike the above methods, in this paper, we designed a new computational model called DuGEL based on the graph convolutional neural network and the graph attention network to infer possible microbe-disease associations. In DuGEL, we first downloaded known microbe-disease associations to form a heterogeneous microbe-disease network. Then, we input this network into a graph convolutional neural network and a graph attention network separately to obtain the local and global features of nodes in the network. Next, we spliced the outputs of the graph convolutional neural network and the graph attention network and then introduced a Long Short-Term Memory (LSTM) network to process the fused features. Finally, the output of the LSTM would be passed to a fully connected layer to infer potential associations between microbes and diseases. Experiments showed that DuGEL obtained satisfactory predictive performance with a 5-fold cross-validated auc of 0.9698 and 0.9119 for HMDAD and Disbiome datasets, respectively, and may be a potential tool for future microbe-disease association prediction.

2 Materials and methods

2.1 Datasets

HMDAD, constructed by Ma et al. (2017), and Disbiome (Janssens et al., 2018), constructed by Janssens et al., are the main publicly available biomedical databases containing microbe-disease association data. As shown in Table 1, HMDAD database covers 483 known microbe-disease associations, and processing these data, we ended up with 450 known microbial-disease associations. The HMDAD database provides a valuable information resource for studying microbial-disease relationships. In addition, the Disbiome, constructed by Janssens et al., is a publicly available database of microbe-disease associations. As shown in Table 1, Disbiome database collects 5,573 known associations from published academic papers for 240 diseases and 1,098 microorganisms. After de-duplication, we had 4,351 known microbe-disease associations covering 218 diseases and 1,052 microorganisms. Due to its extensive data collection and detailed information records, the Disbiome database has become a vital data support for research in this field.

After acquiring the initial data, we performed data preprocessing steps to ensure the quality of the data and the validity of the model training. First, we removed all duplicate records to ensure that the association of each microbe with a disease was unique. Further, we converted the data into a uniform format to facilitate subsequent processing and model training. For simplicity, for each dataset, let $M = \{m_1 m_2, \ldots, m_{N_M}\}$ denote the set of newly downloaded different microorganisms, and $D = \{d_1 d_2, \ldots, d_{N_d}\}$ represent the

set of newly downloaded different diseases. Thus, we can construct a primitive known microbe-disease association network $Net = \langle M \cup D, E \rangle$ as follows: for any given $m_i (1 \leq i \leq N_m)$ and $d_j (1 \leq j \leq N_d)$, if and only if there is a known association between them, we assume that there is an edge belonging to E. Obviously, based on above definition, we can obtain an adjacency matrix $A \in \mathbb{R}^{N_m \times N_d}$ as follows: for any given $m_i (1 \leq i \leq N_m)$ and $d_j (1 \leq j \leq N_d)$ if and only if there is an edge between them in E, there is $A_{i,j} = 1$, otherwise, there is $A_{i,j} = 0$.

2.2 Multiple similarity calculation of disease

2.2.1 Gaussian interaction profile kernel similarity of disease

Based on the assumption that two similar diseases will show similar interaction and non-interaction relationships with the same microorganism, in this section, we adapt the Gaussian interaction profile kernel similarity between a pair of diseases d_i and d_j as follows:

$$GD(d_i, d_j) = exp(-\lambda_d \|A(i,:) - A(j,:)\|^2)$$

Where $A(i,:)$ and $A(j,:)$ represent the ith and j^{th} rows of the adjacency matrix A respectively, and λ_d denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_d = \frac{1}{\left(\frac{1}{n_d}\sum_{i=1}^{n_d}\|A(i,:)\|^2\right)}$$

2.2.2 Cosine similarity of disease

Based on the assumption that if two diseases are similar to each other, then their cosine curves will be more coincident, we introduce the cosine similarity between a pair of diseases d_i and d_j as follows:

$$CD(d_i, d_j) = \frac{A(i,:) \cdot A(j,:)}{|A(i,:)| \times |A(j,:)|}$$

The result of cosine similarity has good stability and certainty, the calculation speed is fast and the result is more intuitive. Suitable for large-scale information retrieval. Where $A(i,:) \cdot A(j,:)$ denotes multiplying the vectors of row i and row j, $|A(i,:)|$ represents the mode of $A(i,:)$, and $|A(j,:)|$ represents the mode of $A(j,:)$. $|A(i,:)| \ast |A(j,:)|$ represents the multiplication of two moduli, and then the vector's product removes the modulus's value. Finally, the cosine value of the angle between the two diseases is obtained, that is the cosine similarity. The calculation result of cosine similarity is between −1 and 1. When the similarity between two diseases is exceptionally high, the calculation result tends to be 1. When the similarity between two diseases is very low, the calculation result tends to −1.

2.2.3 Functional similarity of disease

Based on the assumption that similar diseases tend to interact with similar genes, in this section, we calculate the disease functional similarity based on the functional associations between disease-related genes as follows: Firstly, we download the gene interactions

from the HumanNet database in which, every interaction has an associated log-likelihood score (LLS). And then, for any given diseases d_i and d_j, let $G_i = \{g_{i_1}, g_{i_2}, \ldots, g_{i_m}\}$ and $G_j = \{g_{j_1}, g_{j_2}, \ldots, g_{j_n}\}$ be the gene sets of d_i and d_j separately, we will define the functional similarity between d_i and d_j as follows:

$$DFS(d_i, d_j) = \frac{\sum_{g_k \in G_i} F_{G_j}(g_k) + \sum_{g_k \in G_j} F_{G_i}(g_k)}{m + n}$$

where $F_{G_t}(g_q) = max_{g_p \in G_t}(FSS(g_p, g_q))$ and $FSS(g_p, g_q)$ is the functional similarity score between the genes g_p and g_q, which can be calculated as follows:

$$FSS(g_p, g_q) = \begin{cases} 1 & \text{if } p = q \\ \dfrac{LLS(g_p, g_q) - LLS_{min}}{LLS_{max} - LLS_{min}} & \text{if } p \neq q \end{cases}$$

where LLS_{max} and LLS_{min} represent the maximum value of LLS and the minimum value of LLS in HumanNet, respectively.

Thereafter, by combining the above GIP kernel similarity, disease cosine similarity, and functional similarity of disease, we can obtain an integrated similarity matrix of disease as follows:

$$DS = \frac{GD + CD + DFS}{3}$$

2.3 Multiple similarity calculation of microbe

2.3.1 Gaussian interaction profile kernel similarity of microbe

In the same way, we can calculate the gaussian interaction profile kernel similarity between any two microbes m_i and m_j as follows:

$$MD(m_i, m_j) = exp(-\lambda_m \|A(:,i) - A(:,j)\|^2)$$

where $A(:,i)$ and $A(:,j)$ represent the ith and j^{th} columns of the adjacency matrix A respectively, and λ_m denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_m = \frac{1}{\left(\frac{1}{n_m}\sum_{i=1}^{n_m}\|A(:,i)\|^2\right)}$$

2.3.2 Cosine similarity of microbe

Similarly, the cosine similarity between any two microbes m_i and m_j can be obtained as follows:

$$CM(m_i, m_j) = \frac{A(:,i) \cdot A(:,j)}{|A(:,i)| \times |A(:,j)|}$$

The calculation process of cosine similarity between two microorganisms is the same as that of disease cosine similarity. Similarly, when the similarity between two microorganisms is exceptionally high, the calculation result tends to be 1. When the similarity between two microorganisms is very low, the calculation result tends to −1.

2.3.3 Functional similarity of microbe

The functional similarity of the microbe is calculated by using the following method (Zhang et al., 2018): for any given disease d^t, it
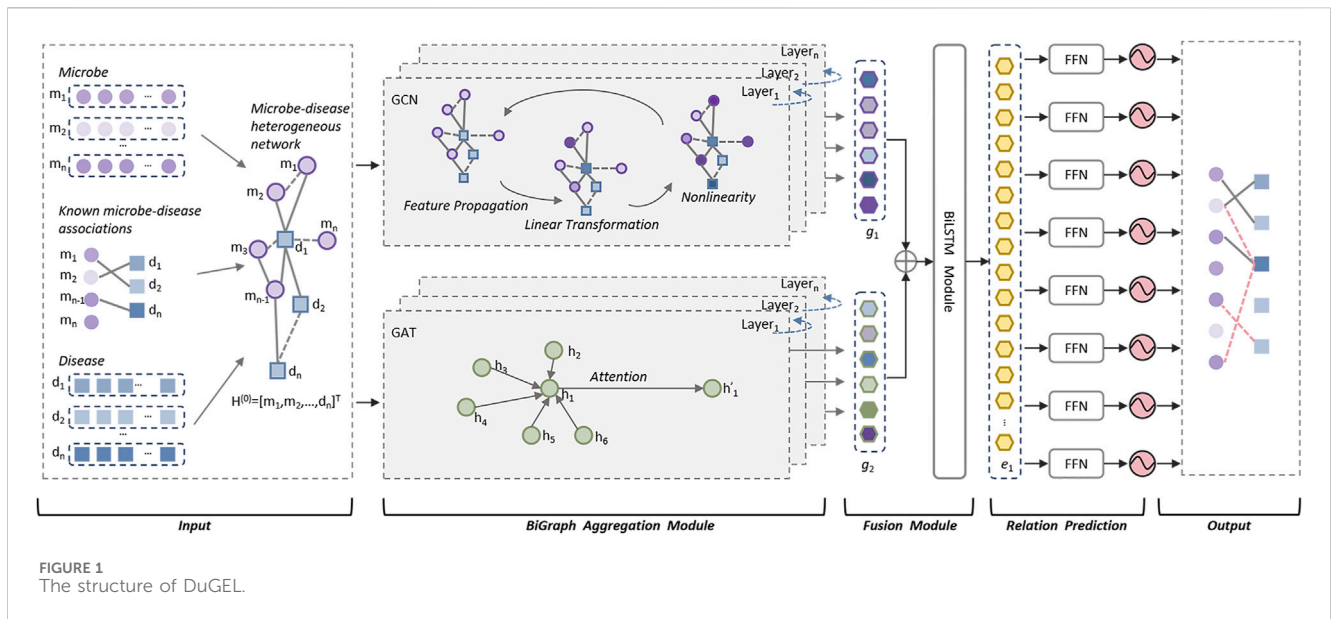
FIGURE 1
The structure of DuGEL.

is first raised Directed Acyclic Graph $DAG_{d_t} = (V_{d_t}, E_{d_t})$, where V_{d_t} includes the disease d_t and its ancestor diseases, E_{d_t} contains all the directed edges from parent nodes to children nodes, and then, the semantic contribution of the disease d_l in V_{d_t} to d_t is defined as:

$$SC_{d_t}(d_l) = \begin{cases} 1 & \text{if } d_l = d_t \\ max\{0.5 \times SC_{d_t}(d_l') \mid d_l' \in \text{children of } d_l\} & \text{otherwise} \end{cases}$$

The semantic value of disease d_t is formulated by:

$$SV_{d_t} = \sum_{d_l \in V_{d_t}} SC_{d_t}(d_l)$$

Then, the semantic similarity between any two diseases d_i and d_j can be defined as follows:

$$DSS(d_i, d_j) = \frac{\sum_{d_i \in V_{d_i} \cap V_{d_j}} \left(SC_{d_i}(d_l) + SC_{d_j}(d_l)\right)}{SV_{d_i} + SV_{d_j}}$$

Relying on the above formulae, we can further define the similarity between the disease d_i and a set of diseases D as follows:

$$DS(d_i, D) = max_{d_j \in D}\left(DSS(d_i, d_j)\right)$$

Hence, for any two given microbes m_i and m_j, we can calculate the function similarity between them as follows:

$$MFS(m_i, m_j) = \frac{\sum_{d_j \in D_i} DS(d_j, D_i) + \sum_{d_j \in D_i} DS(d_j, D_j)}{|D_i| + |D_j|}$$

where D_i denotes the set of diseases associated with the microbe m_i, and D_j represents the set of diseases associated with the microbe m_j. Obviously, by combining the above GIP kernel similarity, disease cosine similarity, and functional similarity of the microbe, we can obtain an integrated similarity matrix of the microbe as follows:

$$MS = \frac{MD + CM + MFS}{3}$$

2.4 Construction of the heterogeneous network

Based on above descriptions, it is easy to see that we can construct a heterogeneous network Y by combining the integrated similarity matrix DS of disease and the integrated similarity matrix MS of microbe with the adjacency matrix A as follows:

$$Y = \begin{bmatrix} DS & A \\ A^T & MS \end{bmatrix}$$

2.5 Structure of the DuGEL

As illustrated in above Figure 1, the DuGEL consists of the following five steps:

- Step 1: Construct a heterogeneous microbe-disease network based on newly downloaded known microbe-disease associations and multiple microbe and disease similarity metrics.
- Step 2: Feeding the heterogeneous microbe-disease network forward into a dual channel structure consisting of a Graph Convolutional Neural Network (GCN) and a Graph Attention Network (GAT), where the GCN is utilized to extract spatial features of nodes in the heterogeneous microbe-disease network from local to global, and the GAT is adopted to assign different importance to the neighbors of each node in the heterogeneous microbe-disease network as it is processed.
- Step 3: Splicing the outputs of GCN and GAT by simply fusing the information captured by GCN and GAT and combining structural and node characteristics and the importance between neighboring nodes.
- Step 4: Implementing a Long Short-Term Memory (LSTM) network to process the fused features, then feeding the output of LSTM into a fully connected layer to convert

the high-level features captured by LSTM into the target output space.

- Step 5: By feeding the newly obtained feature vectors of the target output space into a Sigmoid function for binary prediction, potential associations between microbes and diseases can be finally computed.

2.6 Microbe-disease representation layer

In DuGEL, the input layer is the Microbe-Disease Association Representation Layer, is a component used to convert raw data of known microbe-disease associations into a structured data fame that can be processed by subsequent graph neural networks. Firstly, the newly collected microbe-disease association data need to be pre-processed to ensure consistency and accuracy. The preprocessed data will be used in turn to construct a binary microbe-disease association matrix M, which implies potential relationships between microorganisms and disease (Marsh and Zaura, 2017), and can be defined as follows: Given a microorganism m and a disease n, he known relationship between the microorganism and the disease can be characterized by a binary association matrix $M \in \mathbb{R}^{m \times n}$, where the matrix element M_{ij} is 1 if there is a known association between the microorganism i and the disease j, and 0 otherwise. Each row of the matrix represents a microorganism, and each column represents a disease. The entries in the matrix indicate the presence or absence of an association. The graph structure G is obtained by the association matrix M, where the nodes in the graph are either microbe nodes or disease nodes, and if there is an association between a microbe node and a disease node (Abuin-Denis et al., 2024), i.e., $M_{ij} = 1$, then an edge exists between the nodes.

To enhance the prediction ability of DuGEL, similarity information is also fused in the representation layer (Feng et al., 2023), which contains the microbial similarity matrix S_m and the disease similarity matrix S_d. Among them, the microbial similarity matrix S_m captures similarities between different microorganisms based on genomic, phenotypic, or ecological characteristics. After the microbial similarity matrix S_m and disease similarity matrix S_d are constructed, we fuse this similarity information with the microbe-disease association matrix M to enhance the performance of the association prediction model. Therefore, at this point, we can obtain the association matrix that incorporates microbial similarity and disease similarity as a graph structure (Yu et al., 2024), which further serves as the initial input form for the dual graph-based feature extraction module. This graph structure captures direct associations and enables the model to learn a more comprehensive representation of features.

In addition, we will initialize each representative microbe or disease node in the graph with a feature vector. Formally, let $X_m \in \mathbb{R}^{m \times d}$ represent the initial feature matrix for microbes and $X_d \in \mathbb{R}^{n \times d}$ represent the initial feature matrix for diseases, where d is the dimension of the feature vector. The initial embeddings X_m and X_d are combined into a unified feature matrix $X \in \mathbb{R}^{(m+n) \times d}$. X is fed forward as input to the dual graph-based feature extraction module.

The microbe-disease association representation layer lays the foundation for the entire DuGEL model, and by meticulously structuring the input data, the layer ensures that subsequent graph neural networks are able to effectively capture both local and global patterns in the data. The construction of the graph enables the model to fully utilize all available information, thereby improving the accuracy and robustness of microbe-disease association predictions.

2.7 Bipartite graph feature extraction module

In this study, the dual graph feature extraction module is a core component of the DuGEL designed to extract deep features from the input microbe-disease association feature matrix. The module combines two parallel graph neural network architectures: graph convolutional network (GCN) and graph attention network (GAT), ensuring effective capture of local and global relationships in the graph.

2.7.1 Graph convolution sublayer

The Graph Convolutional Network (GCN) module extracts the spatial features of the graph by processing the microbe-disease feature matrix X to capture the intrinsic structural information of the graph. The spatial features represent the connections between nodes, including direct links (edges) and indirect influences through neighboring nodes. For example, in a microbe-disease graph, the direct association between a microbe and a disease, as well as second- or higher-order neighborhood relationships, contribute to the spatial information. The GCN extracts the features of the nodes by applying convolutional operations on the graph structure, and aggregates the information of the aggregated features of the local neighborhoods of each node by means of the layer-wise propagation rules (Du et al., 2024). The propagation rules of the GCN layers are defined as follows:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right)$$

where $H^{(l)}$ is the node representation X of the l nd layer, $W^{(l)}$ whose initial input is, is the weight matrix of the l-th layer, σ denotes the nonlinear activation function, $\hat{A}$ is the representation of the adjacency matrix A plus the unitary matrix I, and $\hat{D}$ is the corresponding $\hat{A}$ degree matrix.

GCN effectively smoothies the feature representation over the graph structure and ensures that the representation of each node is influenced by its neighbors (Chen et al., 2021), thus capturing the local structure information. By aggregating a node's neighbor information, the feature representation of a node is made to reflect its local graph structure. This aggregation operation is performed in each layer, and through the gradual aggregation of multiple layers (Liu et al., 2018), the GCN can capture a wider range of graph information. This is particularly important for microbe-disease association prediction, as some associations may not be directly visible, but indirectly inferred through multi-hop relationships.

2.7.2 Attention sublayer

The Graph Attention Network (GAT) module introduces an attention mechanism that assigns different importance coefficients to each node's neighbors. For each node i in the graph, GAT computes an attention coefficient $\alpha_{i,j}$ between node i and its neighbor node j, which is learned by a shared attention mechanism:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(a^T\left[W^{(l)}h_i \parallel W^{(l)}h_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)}exp\left(LeakyReLU\left(a^T\left[W^{(l)}h_i \parallel W^{(l)}h_k\right]\right)\right)}$$

where h_i and h_j are the feature vectors of node i and node j, respectively; $W^{(l)}$ is the learnable weight matrix; ∥ represents the splicing operation between the vectors, $\mathcal{N}(i)$ denotes the set of neighboring nodes of node i, and a^T is the weight vector of the attention mechanism. Further, the updated feature vector $h_i^{'}$ of node i is computed by weighted sum of its neighboring features:

$$h_i^{'} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}W^{(l)}h_j\right)$$

The GAT sublayer enables the model to focus on the most relevant parts of the graph, thus capturing the importance of each neighboring node during the feature aggregation process. The advantage is that it can dynamically adjust the contribution of each neighbor to a node's feature update, and by learning different attention weights, GAT can allocate different attention between different parts of the graph structure (Chatzianastasis et al., 2023). This is particularly useful when dealing with complex biological networks, where associations between microorganisms and diseases may have different biological significance and importance.

The dual graph feature extraction module can capture local and global information in the graph structure by combining GCN and GAT approaches. GCN emphasizes the aggregated features of a node's local neighbors. At the same time, GAT dynamically adjusts the weights of the neighboring nodes through the attentional mechanism (Wang et al., 2022), thus providing more flexible and fine-grained control in the feature extraction process. This dual approach ensures that the model understands direct associations between microbes and diseases and identifies potential indirect relationships through graph structural features and attention weights. Combining these two approaches enables the DuGEL model to excel in microbe-disease association prediction tasks, providing more accurate and comprehensive predictions.

The dual graph feature extraction module plays a crucial role in the DuGEL model. Capturing complex graph structure information improves the model's predictive ability and enhances its robustness and generalization ability.

2.8 Feature fusion and sequence learning networks

2.8.1 Fusion layer

After processing in the GCN and GAT layers, we obtain two sets of feature representations $H_{GCN} \in \mathbb{R}^{(m+n)\times d}$ and $H_{GAT} \in \mathbb{R}^{(m+n)\times d}$. These two sets of feature representations capture the spatial properties of the graph and the weighted properties of the important nodes, respectively. For integrating the feature representations generated by the GCN and the GAT, we introduce the fusion layer. These representations encode complementary information: GCN captures spatial structure, while GAT focuses on the relative importance of neighboring nodes.

In the dual graph feature extraction module, we extract two different feature representations through GCN and GAT. The task of the feature fusion layer is to effectively fuse these different features to obtain a comprehensive feature representation. The fusion operation can be realized in various ways, such as concatenation, weighted summing, or multilayer perceptron (MLP) (Afzal et al., 2023). Here, we adopt the concatenation operation to stitch together feature representations from different sub-networks to form a richer feature vector. Assuming that the feature representation extracted through GCN is H_{GCN} and the feature representation extracted through GAT is H_{GAT}, the fused feature representation H_{fusion}:

$$H_{\text{fusion}} = [H_{GCN}||H_{GAT}]$$

where ∥ represents the concatenation operation between the vectors; the fused feature matrix $H_{\text{fusion}} \in \mathbb{R}^{(m+n)\times 2d}$ contains features extracted from two different viewpoints, providing a more comprehensive node representation.

2.8.2 Sequence learning layer

After feature fusion, we need to further learn useful information from these high-dimensional features. The sequence learning layer is designed to capture the temporal or sequential dependencies between features to enhance the prediction capability (Yuan et al., 2023). By treating the node features as a sequence, the order in which node features are fed to the LSTM introduces a dependency chain. The model learns how the features of one node are influenced by or related to those of neighboring nodes. As illustrated in Figure 2, we introduce the Long Short-Term Memory Network (LSTM), which is capable of effectively remembering long-term dependencies and is suitable for processing sequence data (Yoo et al., 2023). The LSTM processes each node's features across multiple time steps. Here, the "time steps" correspond to sequential relationships between node features derived from their embedding in the heterogeneous network. We represent the vector corresponding to the t-th time step in the matrix $H_{\text{fusion}} \in \mathbb{R}^{(m+n)\times 2d}$ as $H_{fusion}^t \in \mathbb{R}^{2d}$. The computational mechanism of the LSTM can be formally represented as follows:

$$I_t = \sigma\left(W_iH_{fusion}^t + U_ih_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_fH_{fusion}^t + U_fh_{t-1} + b_f\right)$$
$$o_t = \sigma\left(W_oH_{fusion}^t + U_oh_{t-1} + b_o\right)$$
$$\tilde{c}_t = tanh\left(W_cH_{fusion}^t + U_ch_{t-1} + b_c\right)$$
$$C_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot tanh\left(c_t\right)$$

where i_t, f_t, and o_t denote the activation vectors of the input, oblivion, and output gates, respectively, $\tilde{c}_t$ is the candidate cell state, c_t and $h_t \in \mathbb{R}^{d_l}$ are the cell state and hidden state at time step, respectively, W and U are the learnable weight matrices, and b is the bias vector. d_l is the dimensionality of the LSTM hidden layer. The output of the LSTM $H_{\text{LSTM}} = \{d_l^t\}_{t=1}^{m+n} \in \mathbb{R}^{(m+n)\times d_l}$ combines information provided by the feature fusion layer, and captures through sequence learning the complex dependencies between features.

The feature fusion layer effectively integrates feature representations extracted from different perspectives, providing more comprehensive input data (Zhang et al., 2020). The sequence learning layer, on the other hand, further enhances the
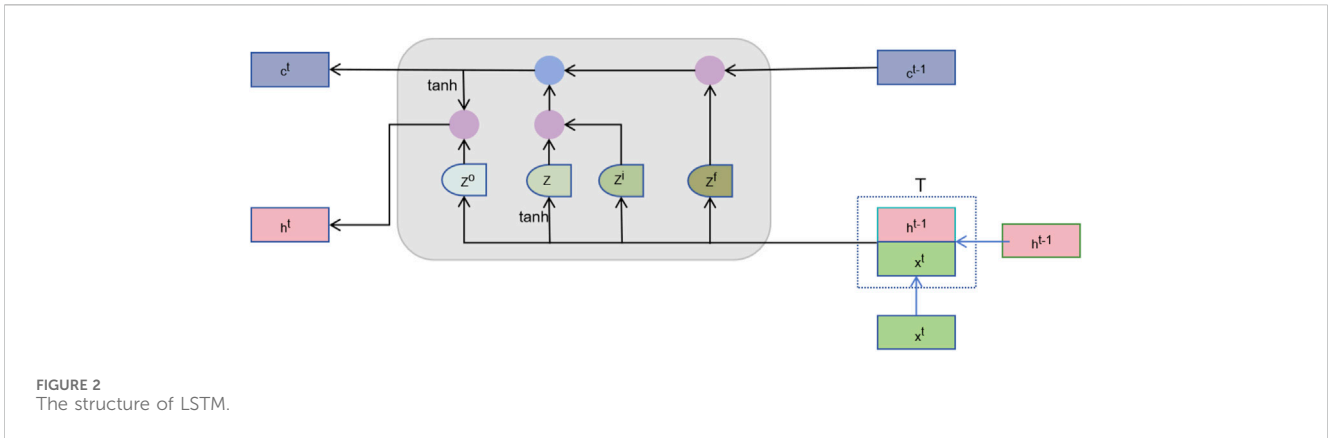
FIGURE 2
The structure of LSTM.

model's predictive capability by capturing the temporal dependencies between features. Combining the two ensures the model can fully utilize all available information to achieve higher accuracy and robustness in microbe-disease association prediction tasks.

2.9 Prediction

The feature representation H_{LSTM} output from the sequence learning layer is used as input to the fully connected layer for further processing. The fully connected layer maps the high-dimensional features to the final prediction results through a series of linear transformations and nonlinear activation functions (Feng et al., 2023). The computational process of the fully connected layer can be formally represented as:

$$\hat{y} = \sigma\left(W_{fc} H_{LSTM} + b_{fc}\right)$$

where W_{fc} and b_{fc} are the weights and bias parameters of the fully connected layer, respectively, and σ represents the Sigmoid nonlinear activation function.

Ultimately, the output layer gives the predicted probability of microbe-disease association (Marsh and Zaura, 2017). By setting an appropriate threshold, it is possible to determine whether there is an association between microbes and diseases.

2.10 Objective function

To measure the difference between the predicted and true values of the model, in the DuGEL model, we use the cross-entropy loss function to evaluate the effectiveness of microbe-disease association prediction (Mao et al., 2023). The cross-entropy loss function is commonly used in classification problems, and in microbe-disease association prediction, we modeled the problem as a binary classification task, i.e., predicting whether there is an association between a certain pair of microorganisms and a disease. The cross-entropy loss function is defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right]$$

where N denotes the number of samples; y_i is the true label of the i-th sample, indicating the presence of an association, and 0 denotes the absence of an association. $\hat{y}_i$ is the predicted probability of the model for the i-th sample, indicating the probability of an association between microorganisms and diseases. The cross-entropy loss function improves the accuracy of the prediction by penalizing the wrong prediction of the model, which makes the model continuously adjust the parameters during the training process.

To prevent model overfitting, we add a regularization term to the loss function. The regularization term improves the generalization ability of the model by adding a penalty to the model complexity in the loss function (Fu et al., 2023), encouraging the model to choose simpler parameter configurations. In the DuGEL model, we use L2 regularization or weight decay. It is defined as follows:

$$\mathcal{L}_{reg} = \sum_{k} \parallel W_k \parallel_2^2$$

where W_k denotes the k rd weight matrix of the model; and $\parallel W_k \parallel_2^2$ is the L2 paradigm of W_k, i.e., the sum of squares of the weight matrix. The regularization term prevents the model from overfitting the training data by penalizing excessively large values of the weights, thus improving the model's performance on the test data.

Ultimately, the integrated loss function of the DuGEL model consists of a cross-entropy loss and a regularization term of the following form:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda * \mathcal{L}_{reg}$$

where λ is the number, which controls the weights of the regularization term. This integrated loss function takes into account both the accuracy of the model prediction and the complexity of the model, and by balancing the two (Cui et al., 2019), it ensures that the model not only accurately fits the training data during training, but also has good generalization ability.

3 Experiments and results

In this section, we will detail the experimental setup, evaluation metrics, and baseline methodology used to evaluate the performance

TABLE 2 Pseudocode of the DuGEL model proposed in this study.

Algorithm 1 Algorithm of the DuGEL.
Input: Known associations matrix A $\in R^{N_m \times N_d}$;

$\hat{A}$: Representation of the adjacency matrix A plus the unitary matrix I
    Microbe similarity matrix $S_m \in R^{N_m \times N_m}$
    Disease similarity matrix $S_d \in R^{N_d \times N_d}$
    N: the number of iterations for DuGEL

L: the number of layers in GCN and GAT
Output: Reconstructed microbe-disease associations matrix A'
1   Phase 1: Construct the heterogeneous network Y $\in R^{(N_m+N_d) \times (N_d \times N_m) N_d \times N_d}$
2      $Y = [S_m, A, A^T, S_d]$
3   Phase 2: Initialize the embedding feature matrix H_0
4   H_0 = Initial embeddings for microbes and diseases
5   Phase 3: for i = 1 to N do
6   for l = 1 to L do
7   Calculate the embedding feature $H_l (l = 1, ..., L)$ in the lth layer according to the GCN formula and GAT:
8            $H_l^{GCN} = \sigma (D^{-0.5} \star \hat{A} \star D^{-0.5} \star H_{l-1} \star W_l)$
9   Update H_l using GAT mechanism:
10   H_l^{GAT} = Attention (H_l)
11   end for
12   end for
13   Phase 4: Feature Fusion and Sequential Learning
14   Concatenate the final embeddings of microbes and diseases:
15   H_{fused} = Concatenate (H_l^{GCN}, H_l^{GAT})
16   Pass H_{fused} through a BiLSTM layer:
17   H_{seq} = BiLSTM(H_{fused})
18   Apply a dense layer to the output of BiLSTM:
19   H_{final} = Dense (H_{seq})
20   Phase 5: Reconstruct Associations
21   Compute the reconstructed association matrix A' using the final embeddings:
22         $A' = \text{sigmoid}(H_{final})$
23   Return A'

of the DuGEL model and present the experimental results and analysis. The effectiveness and superiority of the DuGEL model in the microbe-disease association prediction task are verified by comparing it with multiple baseline methods. The corresponding pseudo-code of the DuGEL model is shown in Table 2.

3.1 Experimental setup

In this study, we extracted disease features, microbial features, and disease-microbe association matrices from the HMDAD and Disbiome databases, and the three feature matrices mentioned above

were used to construct heterogeneous maps to reflect the interactions between diseases and microbes for disease characterization and microbial characterization. The number of training rounds was set to 4,000 in the training phase. To optimize the algorithm to adjust the weights, the learning rate was set to 0.01. We set the random deactivation strategy for the adjacency matrix with the dropout ratio set to 0.5, thus preventing overfitting. For the subject model, to randomly discard some network connections during the training process to enhance the model's generalization ability, we similarly set the random deactivation strategy with the dropout ratio set to 0.5. In addition, we set the similarity weight to 6 to weigh the similarity features of diseases and microorganisms. From the above description, it is easy to see that there are several hyperparameters in DuGEL, such as the dimension k of the embedded features, the number of layers L, the initial learning rate r of the optimizer, the total training epoch α, the node dropout β, and the rule dropout γ. As illustrated in Figures 3, 4, the various results for several combinations of parameters k and L in the 5-fold cv. From the figures, it is easy to see that the optimal combination of k and L is L = 2, k = 128, which indicates that the first and second-layer embedded features contain more information than the third-layer embedded features. After analyzing, this may be due to excessive smoothing of LSTM.

In the training and evaluation phases of the model, we perform multiple cross-validations. In each validation, the dataset is divided into a training set and a test set, which trains the model on the training set and evaluates the model performance on the test set. Specifically, 5-fold cross-validation (k_folds) is used to evaluate the model performance. In each cross-validation, the data are randomly disrupted and divided into five parts, one used as the test set and the rest as the training set. To ensure the stability and reliability of the results, we repeat the execution of the experiment 5 times and report the average performance metrics.

3.2 Evaluation metrics

In order to evaluate the method proposed in this paper, we employ a series of evaluation metrics to comprehensively measure the performance of the model, including AUC, accuracy and specificity The formal definitions and calculations of each evaluation metric are given below:

Accuracy is the ratio of the number of correctly categorized samples to the total number of samples, which is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP denotes True Positives, FP denotes False Positives, FN denotes False Negatives, and TN denotes True Negatives. Accuracy reflects the overall ability of the model to correctly classify.

AUC (Area Under the ROC Curve) represents the area under the receiver operating characteristic curve (ROC Curve), which is used to measure the classification performance of the model. The ROC Curve plots the True Positive Rate (TPR) and False Positive Rate (FPR) through different thresholds. TPR and FPR are defined as follows:
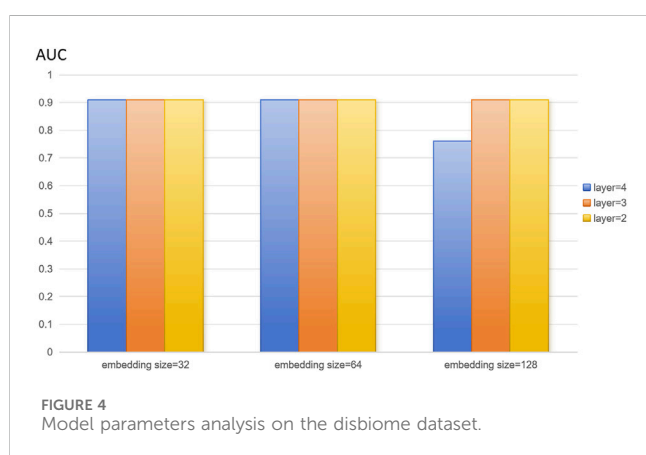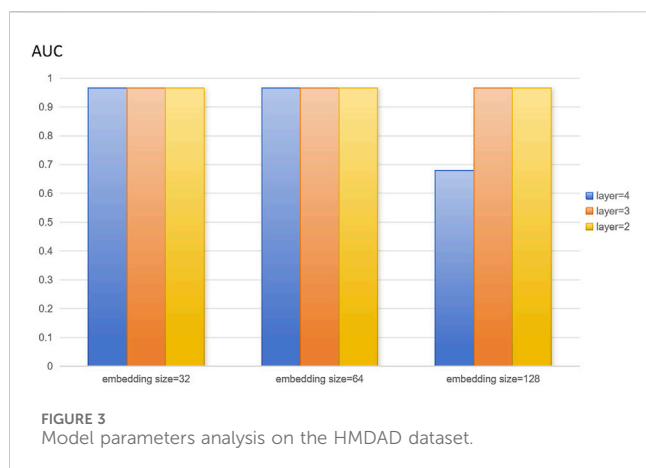
$$TPR(Recall) = \frac{TP}{TP + FN}$$

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

The value of AUC is between 0 and 1, with larger values indicating better model performance.

Specificity, also known as True Negative Rate (TNR), is formally defined as follows:

$$Specificity = \frac{TN}{FP + TN}$$

Specificity indicates the proportion of all samples that are actually negative that are correctly predicted to be negative.

3.3 Alternative methods for microbedisease association prediction

In order to compare the methods proposed in this paper horizontally, as shown in Table 3, we introduced nine microbe-disease association prediction methods of.

The above methods can be categorised into four groups, with network-based methods constructing complex network structures based on known microbial disease associations. These network-

based methods construct complex network structures based on known microbial-disease associations. Then, the potential probability of associations between microorganisms and diseases is inferred by analyzing the structural features of these networks and the lengths and numbers of connecting paths between the nodes (Wang et al., 2011). For example, Chen et al. proposed the KATZHMDA computational model to infer possible microbe-disease associations using the KATZ measure, which takes measurements to capture global information in a network by counting all paths between nodes and then predicts potential associations (Chen et al., 2017). However, the matrix decomposition-based approach focuses on decomposing the known microbe-disease association matrix into two feature matrices and approximating the original association matrix by the product of these two matrices (Ma and Liu, 2022). Information such as similarity and strength of association between microorganisms and diseases can be obtained. Shen et al. proposed a computational model of CMFHMDA based on synergistic matrix decomposition (Shen et al., 2017). In addition, based on the traditional machine learning approach by using known associations as training samples, we can construct a model for predicting the association between unknown microorganisms and diseases (Long et al., 2021). For example, Wang et al. designed the LRLSHMDA model which represents the network structure by constructing a Laplace matrix and predicts the association by regularized least squares (Wang et al., 2017). Finally, the graph neural network-based approach utilizes neural networks to take microbial and disease-related data as inputs and to extract and explore features and patterns from graph-structured data (Bessadok et al., 2022). They can utilize we can utilize the powerful learning ability of neural networks to discover potential associations between microbes and diseases and mine functional patterns and features from complex graph data. For example, Wang et al. designed GCNMA that captures structural information in the network by introducing a graph convolutional neural network and incorporates a multilayer attention mechanism to enhance the ability to model complex relationships between nodes (Wang et al., 2023).

Although the above models can perform reliably in some aspects, they still have limitations. For example, the computation of Katz path correlation must consider all paths between different nodes (Zhang et al., 2017). This may lead to high computational complexity on large datasets, especially when the network size is large (Kumar et al., 2020). In addition, regularized least squares usually introduce a regularization term to avoid overfitting. However, choosing the appropriate regularization parameter is not easy. If it is not chosen correctly, it may lead to underfitting or overfitting problems. When there is noise in the input data, the regularized least squares method may be too sensitive to the noise, resulting in unstable or inaccurate prediction results (Jung and Park, 2017).

3.4 Experimental results and analysis

In this section, we provide a detailed analysis of the experimental results of our proposed DuGEL model on the HMDAD and Disbiome datasets and compare it with nine other state-of-the-art microbe-disease association prediction methods. First, Tables 4, 5

TABLE 3 Microbe-disease association prediction methods.

Method	Approach
KATZHMDA (Chen et al., 2017)	Use of KATZ measurements to infer possible microbe-disease associations
LRLSHMDA (Wang et al., 2017)	Microbial disease prediction using the Laplace regularised least squares framework
NTSHMDA (Luo and Long, 2018)	Predicting potential microbe-disease associations using a model based on stochastic roaming restarts
BiRWMP (Shen et al., 2018)	Predicting microbial-disease using a bidirectional stochastic wandering approach
NBLPIHMDA (Wang et al., 2019)	Detecting potential microbe-disease associations using a bidirectional tag dissemination scheme
HMDApred (Fan et al., 2020)	Predicting microbe-disease associations using network-consistent projections and multiple data integration
BPNNHMDA (Li et al., 2020)	Based on a back propagation neural network design
GATMDA (Long et al., 2021)	Microbial Disease Association Prediction Using Graphical Attention Networks Combined with Matrix Filling
GCNMA (Wang et al., 2023)	Proposing a new computational model based on graph convolutional neural networks and multilayer attention mechanisms

show the performance of our proposed DuGEL model, and the nine compared methods on the HMDAD and Disbiome datasets, respectively, are mainly compared under the AUC assessment metrics. On the HMDAD dataset, the DuGEL model performs well, with its AUC values of 0.9698 and 0.9606 in 5-fold cross-validation and 2-fold cross-validation, respectively, The higher AUC values indicate the more vital overall predictive ability of the model, which indicates that DuGEL can effectively distinguish between positive and negative samples. For the Disbiome dataset, the DuGEL model still performs well. Under 5-fold cross-validation and 2-fold cross-validation settings, DuGEL reaches 0.9119 and 0.8932.

In comparison with previous methods, the DuGEL model demonstrates excellent performance. we attribute the results to the intrinsic structure of DuGEL. DuGEL successfully achieves efficient prediction of microbial-disease associations by combining a graph convolutional neural network (GCN) and a graph Attention Network (GAT), as well as introducing a Long Short-Term Memory Network (LSTM) to process fused features, enabling efficient prediction of microbe-disease associations. This multilevel feature extraction and sequence modeling approach enables DuGEL to perform well on key metrics (e.g., AUPR and AUC),demonstrating its robustness in microbe-disease association prediction tasks.

In particular, the DuGEL model combines the strengths of GCN and GAT; GCN can efficiently capture both local and global spatial features in the graph and extract complex relationships between microbes and diseases by performing convolutional operations on the features of neighboring nodes. Moreover, through the attention mechanism, GAT can assign different importance weights to its neighbors when processing nodes, thus better capturing the information of critical nodes. After combining the two, the outputs of GCN and GAT are fused through the Dual Graph Enhanced Layer of the DuGEL model, which effectively integrates the structural features and node importance.

Furthermore, introducing the LSTM module further enhances the model's capabilities. LSTM is good at processing sequence data and can better identify potential associations between microbes and diseases by capturing temporal dependencies (Baranwal et al., 2022). The memory unit of LSTM can preserve information of long-time dependencies (Wu et al., 2021), which is especially

TABLE 4 HMDAD dataset in 5-fold cv and 2-fold cv.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Chen et al., 2017)	0.8291 ± 0.0041	0.8164 ± 0.0047
LRLSHMDA (Wang et al., 2017)	0.8792 ± 0.0032	0.8589 ± 0.0043
NTSHMDA (Luo and Long, 2018)	0.8893 ± 0.0043	0.8631 ± 0.0052
BiRWMP (Shen et al., 2018)	0.8777 ± 0.0089	0.8693 ± 0.0081
NBLPIHMDA (Wang et al., 2019)	0.8961 ± 0.0033	0.8792 ± 0.0056
HMDApred (Fan et al., 2020)	0.9357 ± 0.0041	0.9049 ± 0.0035
BPNNHMDA (Li et al., 2020)	0.9133 ± 0.0012	0.8949 ± 0.0023
GATMDA (Long et al., 2021)	0.9561 ± 0.0142	0.9541 ± 0.0053
GCNMA (Wang et al., 2023)	0.9610 ± 0.0223	0.9512 ± 0.0076
DuGEL (our model)	0.9698 ± 0.0172	0.9606 ± 0.0057

important for analyzing potential microbe-disease relationships over long periods.

3.5 Ablation experiment

In addition, considering the role and contribution of every sub-module in the DuGEL model proposed in this study, we function ablation experiments on the HMDAD dataset to inspect the effect of distinctive parts of the model. We carried out three specific ablation experiments:

1. Doing away with the diagram convolution sublayer (denoted as "-GCN Layer")
2. Removing the sketch attention sublayer (denoted as "-GAT Layer")
3. Casting off the BiLSTM module (denoted as "-BiLSTM Module")

Overall, each sub-module of the DuGEL model improves the primary mannequin's effectiveness by taking pictures of the

TABLE 5 Disbiome dataset in 5-fold cv and 2-fold cv.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Chen et al., 2017)	0.6781 ± 0.0132	0.6692 ± 0.0063
LRLSHMDA (Wang et al., 2017)	0.7361 ± 0.0221	0.7191 ± 0.0115
NTSHMDA (Luo and Long, 2018)	0.8301 ± 0.0059	0.8079 ± 0.0065
BiRWMP (Shen et al., 2018)	0.8319 ± 0.0092	0.8141 ± 0.0057
NBLPIHMDA (Wang et al., 2019)	0.8429 ± 0.0112	0.8281 ± 0.0142
HMDApred (Fan et al., 2020)	0.8521 ± 0.0381	0.8373 ± 0.0342
BPNNHMDA (Li et al., 2020)	0.8716 ± 0.0191	0.8532 ± 0.0151
GATMDA (Long et al., 2021)	0.9229 ± 0.0081	0.9201 ± 0.0141
GCNMA (Wang et al., 2023)	0.9001 ± 0.0161	0.8803 ± 0.0178
DuGEL (our model)	0.9119 ± 0.0059	0.8932 ± 0.0038

TABLE 6 Fold cv and 2-fold cv and based on HMDAD dataset.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
DuGEL (Proposed)	0.9698 ± 0.0172	0.9606 ± 0.0057
-GCN Layer	0.9465 ± 0.0185	0.9392 ± 0.0046
-GAT Layer	0.9520 ± 0.0163	0.9408 ± 0.0041
-BiLSTM Module	0.9602 ± 0.0053	0.9510 ± 0.0075

TABLE 7 The top 20 Kidney stones microbes predicted by DuGEL.

Rank	Microbe	Evidence
1	*Clostridium difficile*	PMID: 30021693
2	*Helicobacter pylori*	PMID: 25459132
3	*Staphylococcus aureus*	PMID: 14241187
4	*Clostridium* coccoides	PMID: 37609403
5	*Staphylococcus*	PMID: 14241187
6	Bifidobacterium	PMID: 37145061
7	Comamonadaceae	Unconfirmeda
8	Oxalobacteraceae	PMID: 32381601
9	Sphingomonadaceae	PMID: 36970590
10	Dietzia maris	Unconfirmeda
11	*Staphylococcus* epidermidis	PMID: 20466,084
12	*Escherichia coli*	PMID: 14241187
13	*Acinetobacter*	PMID: 32111156
14	Corynebacterium	PMID: 24563271
15	Prevotella copri	PMID: 27708409
16	Propionibacterium	PMID: 33153435
17	Propionibacterium acnes	Unconfirmeda
18	Desulfovibrio	PMID: 38007438
19	Oxalobacter formigenes	PMID: 32880090
20	*Fusobacterium* nucleatum	PMID: 37458823

correlation between microorganisms and diseases. Table 6 suggests the results of the ablation experiments. In the first ablation experiment, after getting rid of the graph convolution sublayer, the mannequin's AUC drops to 0.9465 and 0.9392 in 5-fold cross-validation and 2-fold pass validation, respectively, which suggests that GCN plays a crucial role in shooting nearby and international spatial points between microbes and diseases. By performing convolutional operations on the elements of neighboring nodes, GCN can extract complex correlation information, and removing this component considerably decreases the model's predictive power.

In the second ablation experiment, doing away with the GAT sublayer reduces the model's AUC to 0.9520 and 0.9408 in 5-fold cross-validation and 2-fold pass-validation, respectively. GAT can better capture facts about key nodes via an attention mechanism that assigns unique weights of importance to its neighbors when processing nodes. This mechanism is necessary for improving the model's prediction accuracy, and the mannequin's performance in a similar fashion decreases after its removal.

Finally, we explored the function of the BiLSTM module. After eliminating the BiLSTM module, the AUC of the mannequin diminished to 0.9602 and 0.9510 in the 5-fold move validation and 2-fold move validation, respectively. The LSTM module appropriately processes sequence information and can become aware of attainable associations between microbes and diseases by capturing time dependencies. The reminiscence unit of the LSTM can retain lengthy time-dependent information, which is especially essential for examining the viable microbes' disease. This is mainly necessary for analyzing doable microbe-disease relationships over lengthy periods. The elimination of this module resulted in a considerable reduction in the predictive energy of the model, in addition to demonstrating the necessary function of the BiLSTM module in the DuGEL model.

Overall, every sub-module had a practical impact on the expected performance of the DuGEL model. The GCN and GAT successfully extracted complicated associations between microbes and diseases through shooting spatial features and node importance statistics in the diagram structure, while the BiLSTM module furthermore desirable the predictive capacity of

the mannequin via processing sequence features. The effects of the ablation experiments validate the effectiveness and necessity of these sub-modules and reveal the rationality and superiority of the DuGEL model sketch in the microbe-disease affiliation prediction task.3.6.

4 Case study

In this section, we selected three diseases, kidney stones, eczema, and ileal Crohn's disease, as case studies for HMDAD to validate our model's performance further. Specifically, we ranked these three relevant microorganisms in the prediction score and selected the top 20. Then, we evaluated the predictive performance of DuGEL by searching the literature. Among the

TABLE 8 The top 20 Eczema microbes predicted by DuGEL.

Rank	Microbe	Evidence
1	*Clostridium difficile*	PMID: 27667310
2	*Helicobacter pylori*	PMID: 17568058
3	*Staphylococcus aureus*	PMID: 16965415
4	*Escherichia coli*	PMID: 27667310
5	Dietzia maris	PMID: 26821151
6	*Staphylococcus* epidermidis	PMID: 27416972
7	Stenotrophomonas maltophilia	PMID: 26821151
8	Comamonadaceae	Unconfirmed[a]
9	Oxalobacteraceae	PMID: 32971520
10	Sphingomonadaceae	PMID: 33735474
11	*Acinetobacter*	PMID: 28207943
12	Corynebacterium	PMID: 27562264
13	Prevotella copri	Unconfirmed[a]
14	Oxalobacter formigenes	Unconfirmed[a]
15	Desulfovibrio	PMID: 27812181
16	Tropheryma whipplei	PMID: 2,456,205
17	*Enterococcus*	PMID: 16601353
18	*Pseudomonas*	PMID: 33492004
19	*Staphylococcus*	PMID: 20222931
20	*Clostridium* coccoides	PMID: 24650346

TABLE 9 The top 20 Ileal Crohn's disease microbes predicted by DuGEL.

Rank	Microbe	Evidence
1	Actinobacteria	PMID: 33975420
2	Bacteroidetes	PMID: 32448900
3	*Clostridium* coccoides	PMID: 22719818
4	Firmicutes	PMID: 33975420
5	Prevotella	PMID: 35967326
6	Proteobacteria	PMID: 31530835
7	Enterobacteriaceae	PMID: 36268225
8	Lachnospiraceae	Unconfirmed[a]
9	*Bacteroides* ovatus	PMID: 26275394
10	*Bacteroides* uniformis	PMID: 33102745
11	*Bacteroides* vulgatus	PMID: 26275394
12	Clostridia	PMID: 30818349
13	Faecalibacterium prausnitzii	PMID: 18936492
14	*Clostridium* leptum	PMID: 16188921
15	*Lactobacillus*	PMID: 18839424
16	*Klebsiella*	PMID: 24223596
17	*Staphylococcus*	PMID: 23885156
18	Veillonella	Unconfirmed[a]
19	*Bacteroides*	PMID: 21484962
20	*Escherichia coli*	PMID: 15300573

common diseases, kidney stones cause severe back or abdominal pain accompanied by nausea, vomiting, hematuria, and other symptoms (Stevens, 2018). In recent years, kidney stones have been on the rise. It is prevalent in young adult males (Stamatelou and Goldfarb, 2023). A diverse microbial community exists around renal stones; changes in intestinal and urinary microorganisms may cause the occurrence and development of renal stones. *Clostridium difficile*, Bifidobacterium, and others are more closely associated with kidney stone occurrence (Miller et al., 2022). As shown in Table 7, the relevance of 17 of the top 20 candidate kidney stones-associated microorganisms predicted by DuGEL has been confirmed by previous publications.

In addition, eczema is an inflammatory skin disease. According to many studies, microorganisms are strongly associated with eczema (Zimmermann et al., 2019). People with eczema have a less diverse and less stable skin microbiome than those without. This means the balance of beneficial and harmful bacteria on the skin is disrupted, making it more susceptible to infection and inflammation (Flowers and Grice, 2020). In addition, the presence of *Staphylococcus aureus* is associated with more severe eczema symptoms (Chapsa et al., 2023), suggesting a direct link between this bacteria and the condition. As shown in Table 8 below, it is evident that existing publications have confirmed 17 of the 20 potential eczema-associated microorganisms predicted by DuGEL.

Ileal Crohn's disease is an inflammatory bowel disease. It causes swelling of the tissues of the digestive tract, which may lead to abdominal pain, severe diarrhea, fatigue, weight loss, and malnutrition (Fakhoury et al., 2014). The degree of symptoms ranges from mild to severe and usually comes on gradually, but sometimes, it can come on suddenly without warning. The cause of ileal Crohn's disease is still unknown, but it is often assumed that a virus or bacteria may trigger Crohn's disease. This paper presents a case study of ileal Crohn's disease. As shown in Table 9, it is clear that 18 of the top 20 microorganisms associated with ileal Crohn's disease have been confirmed in the published literature.

5 Discussion and conclusion

Microorganisms play an important role in our lives and exist in countless numbers and diversity. Investigating the potential link between microorganisms and diseases cannot only contribute to the discovery of new therapeutic approaches and preventive strategies but also help advance the field of microbiology and medicine.

In this study, we propose a deep learning model called DuGEL to predict potential microbial disease associations. The DuGEL model combines graph convolutional neural network (GCN), graph attention network (GAT), and long-short-term memory network

(LSTM) to efficiently capture and fuse the complex relationships between microbes and diseases. With the dual-channel structure, DuGEL can extract local and global features in the graph structure and enhance the model's ability to capture critical nodes by assigning different importance weights to neighboring nodes through the attention mechanism. Our comprehensive experiments and case studies consistently show that DuGEL performs very satisfactorily in terms of prediction accuracy.

Although the DuGEL model has been very effective in studying the relationship between microbes and disease, it still has some limitations. Currently, the model relies heavily on the HMDAD and Disbiome datasets. Therefore, future work could focus on expanding the datasets to capture microbe-disease associations more comprehensively and accurately. In addition, the DuGEL model can be applied to drug-target interaction prediction and gene-disease association prediction to validate its broad applicability and effectiveness.

Data availability statement

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

Author contributions

JW: Writing–original draft, Writing–review and editing, Conceptualization, Investigation. LX: Data curation, Formal Analysis, Writing–review and editing. LF: Investigation, Software, Writing–review and editing. LW: Funding acquisition, Methodology, Resources, Writing–review and editing. XZ: Funding acquisition, Supervision, Validation, Writing–review and editing.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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

Abuin-Denis, L., Piloto-Sardiñas, E., Maître, A., Wu-Chuang, A., Mateos-Hernández, L., Obregon, D., et al. (2024). Exploring the impact of Anaplasma phagocytophilum on colonization resistance of Ixodes scapularis microbiota using network node manipulation. *Curr. Res. Parasitol. and Vector-Borne Dis.* 5, 100177. doi:10.1016/j.crpvbd.2024.100177

Afshari, S. S., Enayatollahi, F., Xu, X., and Liang, X. (2022). Machine learning-based methods in structural reliability analysis: a review. *Reliab. Eng. and Syst. Saf.* 219, 108223. doi:10.1016/j.ress.2021.108223

Afzal, S., Ziapour, B. M., Shokri, A., Shakibi, H., and Sobhani, B. (2023). Building energy consumption prediction using multilayer perceptron neural network-assisted models; comparison of different optimization algorithms. *Energy* 282, 128446. doi:10.1016/j.energy.2023.128446

Anisman, H., Hayley, S., and Kusnecov, A. W. (2018). *The immune system and mental health.* Academic Press. doi:10.1111/cei.13334

Baranwal, M., Clark, R. L., Thompson, J., Sun, Z., Hero, A. O., and Venturelli, O. S. (2022). Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics. *Elife* 11, e73870. doi:10.7554/eLife.73870

Barr, J. J. (2017). A bacteriophages journey through the human body. *Immunol. Rev.* 279 (1), 106–122. doi:10.1111/imr.12565

Bessadok, A., Mahjoub, M. A., and Rekik, I. (2022). Graph neural networks in network neuroscience. *IEEE Trans. Pattern Analysis Mach. Intell.* 45 (5), 5833–5848. doi:10.1109/tpami.2022.3209686

Bocci, V. (1992). The neglected organ: bacterial flora has a crucial immunostimulatory role. *Perspect. Biol. Med.* 35 (2), 251–260. doi:10.1353/pbm.1992.0004

Burr, A. H., Bhattacharjee, A., and Hand, T. W. (2020). Nutritional modulation of the microbiome and immune response. *J. Immunol.* 205 (6), 1479–1487. doi:10.4049/jimmunol.2000419

Chapsa, M., Rönsch, H., Löwe, T., Gunzer, F., Beissert, S., and Bauer, A. (2023). The role of bacterial colonisation in severity, symptoms and aetiology of hand eczema: the importance of *Staphylococcus aureus* and presence of commensal skin flora. *Contact Dermat.* 89 (4), 270–276. doi:10.1111/cod.14384

Chatzianastasis, M., Lutzeyer, J., Dasoulas, G., and Vazirgiannis, M. (2023). "Graph ordering attention networks," in Proceedings of the AAAI Conference on Artificial Intelligence, USA, February 2023, 7006–7014. doi:10.1609/aaai.v37i6.25856

Chen, W., Feng, F., Wang, Q., He, X., Song, C., Ling, G., et al. (2021). Catgcn: graph convolutional networks with categorical node features, *IEEE Trans. Knowl. Data Eng.* 35(4): 3500–3511. doi:10.1016/j.apcatb.2018.01.024

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 (5), 733–739. doi:10.1093/bioinformatics/btw715

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). "Class-balanced loss based on effective number of samples," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, USA, June 15th, 2025, 9268–9277. doi:10.1109/CVPR.2019.00949

Du, H., Yu, R., Bai, L., and Wang, W. (2024). Learning structure perception MLPs on graphs: a layer-wise graph knowledge distillation framework, *Int. J. Mach. Learn. Cybern.* 15 4357–4372. doi:10.1007/s13042-024-02150-2

Fakhoury, M., Negrulj, R., Mooranian, A., and Al-Salami, H. (2014). Inflammatory bowel disease: clinical aspects and treatments. *J. Inflamm. Res.* 7, 113–120. doi:10.2147/JIR.S65979

Fan, Y., Chen, M., Zhu, Q., and Wang, W. (2020). Inferring disease-associated microbes based on multi-data integration and network consistency projection. *Front. Bioeng. Biotechnol.* 8, 831. doi:10.3389/fbioe.2020.00831

Feng, S., Zhao, L., Shi, H., Wang, M., Shen, S., and Wang, W. (2023). One-dimensional VGGNet for high-dimensional data. *Appl. Soft Comput.* 135, 110035. doi:10.1016/j.asoc.2023.110035

Flowers, L., and Grice, E. A. (2020). The skin microbiota: balancing risk and reward. *Cell. host and microbe* 28 (2), 190–200. doi:10.1016/j.chom.2020.06.017

Fu, S., Tian, Y., Tang, J., and Liu, X. (2023). Cost-sensitive learning with modified Stein loss function. *Neurocomputing* 525, 57–75. doi:10.1016/j.neucom.2023.01.052

Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., et al. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535 (7610), 94–103. doi:10.1038/nature18850

Grasso, L. L., Martino, D. C., and Alduina, R. (2016). Production of antibacterial compounds from Actinomycetes. *Actinobacteria-basics Biotechnol. Appl.* 7, 177–198. doi:10.5772/61525

Heintz-Buschart, A., and Wilmes, P. (2018). Human gut microbiome: function matters. *Trends Microbiol.* 26 (7), 563–574. doi:10.1016/j.tim.2017.11.002

Hoffmann, A. R., Proctor, L., Surette, M., and Suchodolski, J. (2016). The microbiome: the trillions of microorganisms that maintain health and cause disease in humans and companion animals, *Veterinary pathol.* 53(1): 10–21. doi:10.1177/0300985815595517

Islam, M. A., Haque, M. A., Rahman, M. A., Hossen, F., Reza, M., Barua, A., et al. (2022). A review on measures to rejuvenate immune system: Natural mode of protection against coronavirus infection. *Front. Immunol.* 13, 837290. doi:10.3389/fimmu.2022.837290

Jagannathan, S. V., Manemann, E. M., Rowe, S. E., Callender, M. C., and Soto, W. (2021). Marine actinomycetes, new sources of biotechnological products. *Mar. Drugs* 19 (7), 365. doi:10.3390/md19070365

Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., et al. (2018). Disbiome database: linking the microbiome to disease. *BMC Microbiol.* 18, 50–56. doi:10.1186/s12866-018-1197-5

Jin, D., Song, X., Yu, Z., Liu, Z., Zhang, H., Cheng, Z., et al. (2021). "Bite-gcn: a new GCN architecture via bidirectional convolution of topology and features on text-rich networks," in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, China, 12 October 2021, 157–165. doi:10.1145/3437963.3441774

Jin, H., Hu, G., Sun, C., Duan, Y., Zhang, Z., Liu, Z., et al. (2022). mBodyMap: a curated database for microbes across human body and their associations with health and diseases. *Nucleic Acids Res.* 50 (D1), D808–D816. doi:10.1093/nar/gkab973

Jung, S. M., and Park, P. (2017). Stabilization of a bias-compensated normalized least-mean-square algorithm for noisy inputs. *IEEE Trans. Signal Process.* 65 (11), 2949–2961. doi:10.1109/TSP.2017.2675865

Kumar, A., Mishra, S., Singh, S. S., Singh, K., and Biswas, B. (2020). Link prediction in complex networks based on significance of higher-order path index (SHOPI). *Phys. A Stat. Mech. its Appl.* 545, 123790. doi:10.1016/j.physa.2020.124289

Li, H., Wang, Y., Zhen, Z. Z., Tan, Y., and Wang, L. (2020). Identifying microbe-disease association based on a novel back-propagation neural network model. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* PP (99), 1. doi:10.1109/TCBB.2020.2986459

Li, X., Sun, L., Ling, M., and Peng, Y. (2023). A survey of graph neural network based recommendation in social networks. *Neurocomputing* 549, 126441. doi:10.1016/j.neucom.2023.126441

Liang, D., Leung, R.K.-K., Guan, W., and Au, W. W. (2018). Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathog.* 10, 3–9. doi:10.1186/s13099-018-0230-4

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). "Path aggregation network for instance segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, USA, 18-23 June 2018, 8759–8768. doi:10.1109/CVPR.2018.00913

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Briefings Bioinforma.* 22 (3), 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. Bioinforma.* 17 (4), 1341–1351. doi:10.1109/TCBB.2018.2883041

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe–disease associations. *Briefings Bioinforma.* 18 (1), 85–97. doi:10.1093/bib/bbw005

Ma, Y., and Liu, Q. (2022). Generalized matrix factorization based on weighted hypergraph learning for microbe-drug association prediction. *Comput. Biol. Med.* 145, 105503. doi:10.1016/j.compbiomed.2022.105503

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd_Allah, E. F. (2019). Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Front. Immunol.* 9, 2868. doi:10.3389/fimmu.2018.02868

Mao, A., Mohri, M., and Zhong, Y. (2023). "Cross-entropy loss functions: theoretical analysis and applications," in International Conference on Machine Learning, China, 14 Apr 2023 (PMLR), 23803–23828. doi:10.48550/ARXIV.2304.07288

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12, 634511. doi:10.3389/fmicb.2021.634511

Marsh, P., and Zaura, E. (2017). Dental biofilm: ecological interactions in health and disease. *J. Clin. periodontology* 44, S12–S22. doi:10.1111/jcpe.12679

Miller, A. W., Penniston, K. L., Fitzpatrick, K., Agudelo, J., Tasian, G., and Lange, D. (2022). Mechanisms of the intestinal and urinary microbiome in kidney stone disease. *Nat. Rev. Urol.* 19 (12), 695–707. doi:10.1038/s41585-022-00647-5

Montaner, J., Ramiro, L., Simats, A., Tiedt, S., Makris, K., Jickling, G. C., et al. (2020). Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat. Rev. Neurol.* 16 (5), 247–264. doi:10.1038/s41582-020-0350-6

Munikoti, S., Agarwal, D., Das, L., Halappanavar, M., and Natarajan, B. (2023). Challenges and opportunities in deep reinforcement learning with graph neural networks: a comprehensive review of algorithms and applications. *IEEE Trans. neural Netw. Learn. Syst.* 35, 15051–15071. doi:10.1109/tnnls.2023.3283523

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. big data* 2, 1–21. doi:10.1186/s40537-014-0007-7

Pickard, J. M., Zeng, M. Y., Caruso, R., and Núñez, G. (2017). Gut microbiota: role in pathogen colonization, immune responses, and inflammatory disease, *Immunol. Rev.* 279(1): 70–89. doi:10.1111/imr.12567

Sanz, Y., Olivares, M., Moya-Pérez, Á., and Agostoni, C. (2015). Understanding the role of gut microbiome in metabolic disease risk. *Pediatr. Res.* 77 (1), 236–244. doi:10.1038/pr.2014.170

Shen, D., Liu, J., Wu, Z., Yang, J., and Xiao, L. (2021). ADMM-HFNet: a matrix decomposition-based deep approach for hyperspectral image fusion. *IEEE Trans. Geoscience Remote Sens.* 60, 1–17. doi:10.1109/TGRS.2021.3112181

Shen, X., Zhu, H., Jiang, X., Hu, X., and Yang, J. (2018). "A novel approach based on bi-random walk to predict microbe-disease associations," Intelligent Computing Methodologies: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018. Springer, 746–752. doi:10.1039/C8EE02656D

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 (Springer), 261–269. doi:10.1007/978-3-319-63312-1_24

Stamatelou, K., and Goldfarb, D. S. (2023). Epidemiology of kidney stones, Healthcare. *MDPI*, 424. doi:10.3390/healthcare11030424

Stevens, S. (2018). Obstructive kidney disease. *Nurs. Clin.* 53 (4), 569–578. doi:10.1016/j.cnur.2018.07.007

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 (1), 7601. doi:10.1038/s41598-017-08127-2

Wang, J., Mo, H., Wang, F., and Jin, F. (2011). Exploring the network structure and nodal centrality of China's air transport network: a complex network approach. *J. Transp. Geogr.* 19 (4), 712–721. doi:10.1016/j.jtrangeo.2010.08.012

Wang, L., Wang, Y., Li, H., Feng, X., Yuan, D., and Yang, J. (2019). A bidirectional label propagation based computational model for potential microbe-disease association prediction. *Front. Microbiol.* 10, 684. doi:10.3389/fmicb.2019.00684

Wang, L., Yang, X., Kuang, L., Zhang, Z., Zeng, B., and Chen, Z. (2023). Graph convolutional neural network with multi-layer attention mechanism for predicting potential microbe-disease associations. *Curr. Bioinforma.* 18 (6), 497–508. doi:10.2174/1574893618666230316113621

Wang, M., Wu, L., Li, M., Wu, D., Shi, X., and Ma, C. (2022). Meta-learning based spatial-temporal graph attention network for traffic signal control. *Knowledge-based Syst.* 250, 109166. doi:10.1016/j.knosys.2022.109166

Wu, K., Wu, J., Feng, L., Yang, B., Liang, R., Yang, S., et al. (2021). An attention-based CNN-LSTM-BiLSTM model for short-term electric load forecasting in integrated energy system. *Int. Trans. Electr. Energy Syst.* 31 (1), e12637. doi:10.1002/2050-7038.12637

Wu, Z., Li, W., Liu, G., and Tang, Y. (2018). Network-based methods for prediction of drug-target interactions. *Front. Pharmacol.* 9, 1134. doi:10.3389/fphar.2018.01134

Yoo, G., Kim, H., and Hong, S. (2023). Prediction of cognitive load from electroencephalography signals using long short-term memory network. *Bioengineering* 10 (3), 361. doi:10.3390/bioengineering10030361

Yu, S., Wang, H., Hua, M., Liang, C., and Sun, Y. (2024). Sparse graph cascade multi-kernel fusion contrastive learning for microbe–disease association prediction. *Expert Syst. Appl.* 252, 124092. doi:10.1016/j.eswa.2024.124092

Yuan, X., Huang, L., Li, L., Wang, K., Wang, Y., Ye, L., et al. (2023). Multiscale dynamic feature learning for quality prediction based on hierarchical sequential generative network. *IEEE Sensors J.* 23, 19561–19570. doi:10.1109/jsen.2023.3290163

Zhang, C., Yang, Z., He, X., and Deng, L. (2020). Multimodal intelligence: representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* 14 (3), 478–493. doi:10.1109/JSTSP.2020.2987728

Zhang, W., Yang, W., Lu, X., Huang, F., and Luo, F. (2018). The bi-direction similarity integration method for predicting microbe-disease associations. *Ieee Access2018* 6, 38052–38061. doi:10.1109/ACCESS.2018.2851751

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2017). KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 16 (2), 407–416. doi:10.1109/tcbb.2017.2704587

Zhou, X., Zhang, T., Cheng, C., and Song, S. (2023). Dynamic multichannel fusion mechanism based on a graph attention network and BERT for aspect-based sentiment classification. *Appl. Intell.* 53 (6), 6800–6813. doi:10.1007/s10489-022-03851-3

Zhu, B., Wang, X., and Li, L. (2010). Human gut microbiome: the second genome of human body. *Protein and Cell.* 1 (8), 718–725. doi:10.1007/s13238-010-0093-z

Zimmermann, P., Messina, N., Mohn, W. W., Finlay, B. B., and Curtis, N. (2019). Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: a systematic review. *J. Allergy Clin. Immunol.* 143 (2), 467–485. doi:10.1016/j.jaci.2018.09.025