



A Graph-enhanced Contrastive Learning Framework With Structural and Semantic Representation Learning for Microbe-disease Association Prediction*

SHI Kai^{1,2,3)**}, XU You-Ang¹⁾, HUANG Kai¹⁾, JI Qing-Rong¹⁾, CAO Rui¹⁾,
 WANG Yan-Jun¹⁾, JIANG Qiong^{4)**}

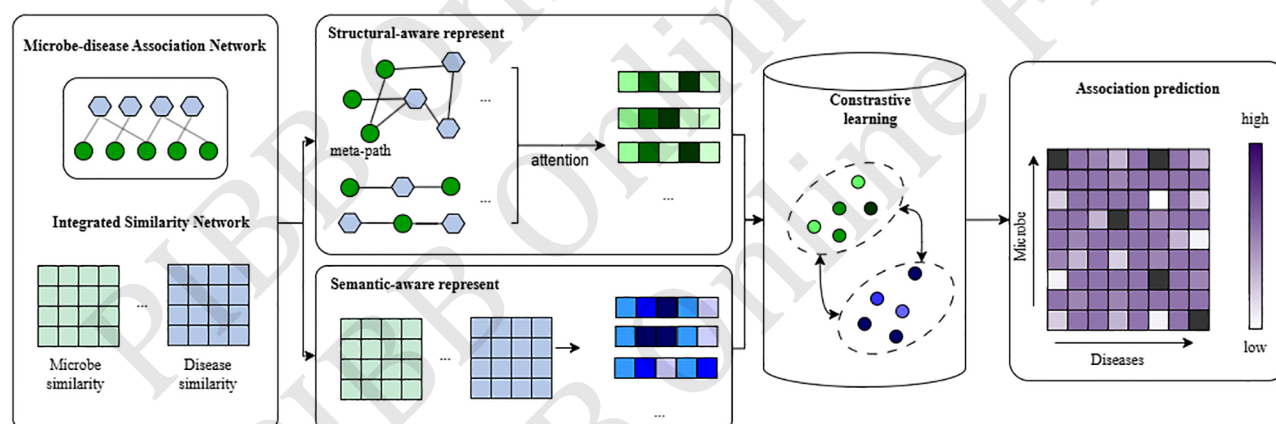
¹⁾College of Computer Science and Engineering, Guilin University of Technology, Guilin 541004, China;

²⁾Guangxi Key Laboratory of Embedded Technology and Intelligent Systems, Guilin University of Technology, Guilin 541004, China;

³⁾Guangxi Academy of Artificial Intelligence, Nanning 530028, China;

⁴⁾College of Physics and Electronic Information Engineering, Guilin University of Technology, Guilin 541004, China)

Graphical abstract



Abstract Objective Gut microbiota dysbiosis has been implicated in various diseases, making the identification of microbe - disease associations essential for understanding disease mechanisms and discovering disease-associated microbes. Computational approaches have been widely used to predict microbe-disease associations and prioritize promising candidates for experimental validation. However, existing methods often suffer from insufficient utilization of heterogeneous biological information, particularly the complementary structural and semantic information derived from association networks and similarity data. Moreover, the sparsity, noise, and incompleteness of known associations further limit prediction performance and generalizability. **Methods** that primarily rely on a single information source or simple feature fusion may fail to capture the complex relationships between microbes and diseases. In this paper, we propose a graph-enhanced contrastive learning framework for microbe - disease association

* This work was supported by grants from The National Natural Science Foundation of China (62562022), and Guangxi Natural Science Foundation (2025JJA170175).

** Corresponding author.

SHI Kai. Tel: 86-13737707647, E-mail: mail_shikai@foxmail.com

JIANG Qiong. Tel: 86-773-5896435, E-mail: huichao1228@163.com;

Received: June 15, 2026 Accepted: October 17, 2026

prediction, GECLMDA, which integrates complementary structural and semantic information to obtain informative node representations and improve the prediction of potential associations. **Methods** Our method derives informative node representations by integrating structural and semantic learning. Specifically, graph convolutional networks are employed to capture structural relationships between microbes and diseases, while a deep sparse autoencoder is used to extract latent semantic features from similarity information. The structural learning component captures relational information from graph-based representations, whereas the semantic learning component extracts complementary information embedded in microbial and disease similarities. Unlike methods that directly contrast multiple graph-derived representations, GECLMDA integrates graph-based relational information before performing cross-representation alignment. Contrastive learning is then introduced to align the structural and semantic views, allowing the model to jointly exploit relational structures and latent similarity patterns. Finally, the resulting embeddings are fused and used to predict potential microbe – disease associations. The method is evaluated on three benchmark datasets, HMDAD, Disbiome, and Peryton, using five-fold cross-validation under different evaluation settings. Ablation experiments and case studies are further conducted to assess the contributions of key model components and the ability of GECLMDA to identify potential disease-associated microbes. In addition, ablation experiments are performed to investigate the contributions of the major model components, and case studies are conducted to examine the ability to identify potential disease-associated microbes. **Results** Our method achieves competitive performance against most baseline methods across the evaluated settings. Under global five-fold cross-validation, GECLMDA obtains *AUCs* of 0.980 ± 0.006 9, 0.944 ± 0.008 3, and 0.954 ± 0.006 7 on HMDAD, Disbiome, and Peryton, respectively. Ablation experiments indicate that structural representation learning, semantic representation learning, and contrastive alignment provide complementary contributions to prediction performance. Case studies further suggest that our method can serve as a potential tool for predicting disease-associated microbes. **Conclusion** The proposed framework effectively integrates structural and semantic information through graph-enhanced contrastive learning and achieves competitive prediction performance across multiple benchmark datasets. GECLMDA provides a potential computational tool for microbe – disease association prediction and the identification of disease-associated microbes, offering useful candidates for further biological investigation.

Key words microbe-disease association, association prediction, meta-path, graph convolutional network, deep sparse autoencoder, contrastive learning

DOI: 10.3724/j.pibb.2026.0325

CSTR: 32369.14.pibb.20260325

The human body harbors diverse microbial communities in the gut, oral cavity, skin, and respiratory tract, where they form a complex microecological system with the host^[1-2]. These microbial communities contribute to digestion and vitamin synthesis, and also play important roles in immune development, metabolic homeostasis, and resistance to pathogen colonization^[3-4]. Growing evidence indicates that gut microbial dysbiosis is more than a consequence of disease and may contribute to the onset, progression, and persistence of various disorders, sometimes as a driving factor^[5-6]. Systematic identification of microbe – disease associations is important for understanding disease mechanisms and identifying potential intervention targets^[7]. However, experimental validation is often costly and time-consuming, which limits its use in large-scale and systematic screening. Computational approaches can efficiently prioritize candidate associations from existing biomedical data and provide useful candidates for further experimental validation.

Various computational models have been developed for microbe – disease association prediction in recent years. These models can generally be grouped into three categories: network-based methods, matrix factorization-based methods, and machine learning-based methods^[8-11]. Network-based

methods typically construct heterogeneous networks by integrating microbe similarity, disease similarity, and known associations, and then infer potential links *via* random walk, path search, or label propagation^[12]. Subsequent studies further improved prediction performance by incorporating multi-similarity fusion and label propagation strategies^[13-15]. However, these methods depend strongly on network structure. Their performance is often limited when prior association information is scarce, especially for new diseases or new microbes. Matrix factorization-based methods usually treat microbe – disease association prediction as a matrix completion problem and recover unknown associations by learning low-dimensional latent representations. Previous studies have shown that non-negative matrix factorization^[16] and logistic matrix factorization with neighborhood regularization^[17] can achieve good performance in this task. These methods handle sparse data effectively by exploiting the intrinsic low-rank structure of the association matrix, yet their optimization cost grows with data scale and the learned latent factors offer limited biological interpretability.

Machine learning methods, particularly deep learning methods, have also been widely applied in microbe-disease association prediction^[18]. Unlike traditional machine learning methods that rely on handcrafted features together with classifiers such as

adaptive boosting^[19], deep models can automatically extract hierarchical representations from raw similarity and interaction data. Early deep learning efforts employed autoencoders and multi-layer perceptrons for feature learning. More recently, graph neural networks have attracted considerable attention owing to their natural suitability for modeling relational data. Methods such as graph convolutional networks (GCDSAEMDA^[20]), graph attention networks, and variational autoencoders (DVAMDA^[21]), have been proposed to capture both local neighborhood information and global structural patterns. Furthermore, contrastive learning, multi-view learning, and attention-based fusion mechanisms have been introduced to enhance representation quality by integrating complementary signals from heterogeneous sources.^[9, 22-26] These methods are effective in modeling complex nonlinear relationships and higher-order structural dependencies, and therefore often achieve strong predictive performance^[27-28].

Despite these advances, microbe - disease association prediction still faces several important challenges. Existing models often do not make full use of heterogeneous biological information, especially the joint modeling of structural relationship information and biological similarity information, which limits their ability to capture complex association patterns. In addition, data noise and incomplete association information can easily affect model performance, resulting in unstable predictions and poor generalization.

Motivated by the previous studies^[9, 29-30], we propose a graph-enhanced contrastive learning framework (GECLMDA) to integrate complementary structural and semantic representations within a unified framework for microbe-disease association prediction. Given the integrated similarity profiles and the microbe - disease association network, the framework learns node representations through two complementary branches. The structural branch focuses on modeling interaction patterns through node-level and meta-path-based graph convolution, whereas the semantic branch captures latent semantic characteristics from integrated similarity features using a deep sparse autoencoder. The representations learned from these two branches are then aligned through contrastive learning and used for downstream association prediction. Results from multiple benchmark datasets and evaluation settings show that GECLMDA achieves competitive performance against most existing state-of-the-art methods. Case studies further show that the proposed method can

identify potentially meaningful microbe - disease associations that are supported by biological evidence.

1 Material and methods

1.1 The known human microbe-disease associations

In current, several databases have been established to catalog known microbe-disease associations, including: HMDAD^[31], Disbiome^[32], gutMDisorder^[33], peryton^[34]. For model evaluation, we select HMDAD, which holds 450 non-redundant associations among 292 microbes and 39 diseases, and Disbiome, with 8 645 associations involving 351 microbes and 1 582 diseases, offering a wider validation. Peryton, contains 1 747 associations between 469 microbes and 38 diseases, is an additional dataset used for global five-fold cross-validation. The statistics about these datasets are displayed in Table 1.

In each dataset, the relationships between microbes and diseases are represented by an adjacency matrix, $A \in R_{n_d \times n_m}$, if there is an association between diseases d_i and microbes m_j , $A_{ij} = 1$, otherwise, $A_{ij} = 0$.

Table 1 The statistics for microbe-disease association datasets

Datasets	Microbe	Disease	Associations
HMDAD	292	39	450
Disbiome	351	1 582	8 645
Peryton	469	38	1 747

1.2 Construction of microbe and disease similarity network

1.2.1 Calculation and fusion similarity

In our previous work^[9], we comprehensively computed similarities for microbes and diseases from two complementary perspectives: network topology and biological function. Topological similarities, derived directly from the known microbe-disease adjacency matrix, include Gaussian Interaction Profile (GIP) kernel similarity, Cosine similarity, and Jaccard similarity. Functional similarities integrate multi-source biological data: for diseases, these encompass symptom-based similarity, semantic similarity (MeSH-based), and gene functional similarity; for microbes, they include sequence similarity, evolutionary distance similarity, and disease semantic-based functional similarity. Further details of the preprocessing are provided in our previous work^[9].

To enhance robustness and capture

complementary information, Similarity Network Fusion (SNF) was applied to non-linearly integrate the multiple similarity measures within each perspective (topology or function), yielding fused topological matrices (DS_{top} , MS_{top}) and fused functional matrices (DS_{fun} , MS_{fun}). Finally, a weighted linear integration combines these two fused perspectives to generate consolidated similarity profiles:

$$\begin{aligned} D_s &= \mu \cdot DS_{top} + (1 - \mu) \cdot DS_{fun}, \\ M_s &= \mu \cdot MS_{top} + (1 - \mu) \cdot MS_{fun} \end{aligned} \quad (1)$$

Here, μ is a tunable weighting parameter. The resulting matrices D_s and M_s serve as the foundational similarity features for constructing the heterogeneous network and for subsequent representation learning in our contrastive framework.

1.2.2 Construction of heterogeneous networks based on meta-path

The microbe-disease association network is a

heterogeneous information network (HIN), capturing complex structural and semantic relationships. To extract high-level relational features, we employ meta-path-based strategies, where a meta-path denotes a composite relation defined by a sequence of entity types and edge types. In this study, we define the meta-path for microbes $MP^m = M - D - M$, which connects microbes sharing associations with the same disease. Similarly, for diseases, the meta-path $MP^d = D - M - D$ connects diseases linked via common microbes.

$$\begin{aligned} A_{MP^m} &= A \times A^T \\ A_{MP^d} &= A^T \times A \end{aligned} \quad (2)$$

where A is the adjacency matrix encoding the associations between microbes and diseases, t indicates the network type (microbial or disease), m and d refer to microbe and disease respectively. An example for constructing the meta-path network has been represented in Figure 1.

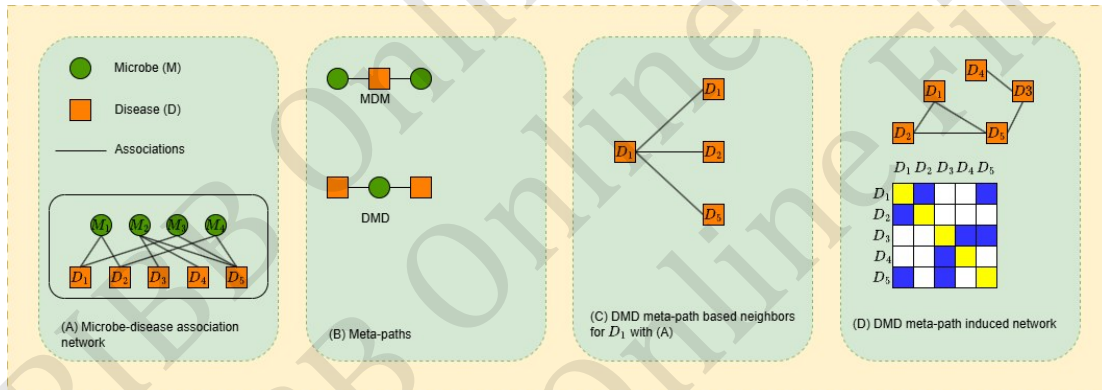


Fig. 1 A meta-path example for GECLMDA

(a) Microbe-disease association network. (b) Four meta-paths involved in GECLMDA, which are MDM, DMD. (c) Disease D_1 and its meta-path-based neighbors D_2, D_3 and D_5 based on the microbe-disease association network in (a). (d) The meta-path-induced network with DMD based on the network in (a).

1.3 GECLMDA

In this study, GECLMDA is a graph-enhanced contrastive learning framework for microbe - disease association prediction (Figure 2). It comprises two complementary branches: the structural branch captures interaction features using node-level and meta-path-based graph convolutions, whereas the semantic branch extracts latent semantic features from the integrated similarity profiles using a deep sparse autoencoder. The resulting representations are subsequently aligned through contrastive learning and jointly used for microbe-disease association prediction.

1.3.1 Structural-aware representation learning

1.3.1.1 Node-level graph convolution on similarity networks

Graph convolutional networks have exhibited expressive capability in learning node embeddings from graph structures. For a graph with symmetric adjacency matrix A , a single-layer graph convolutional encoder is defined as:

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{Q} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \quad (3)$$

Where $\sigma(\cdot)$ is the activation function. $\tilde{Q} = Q + I$ and I is the identity matrix with the same shape as Q , $\tilde{D}$ is the degree matrix of $\tilde{Q}$, $W^{(l)}$ is the learnable weights

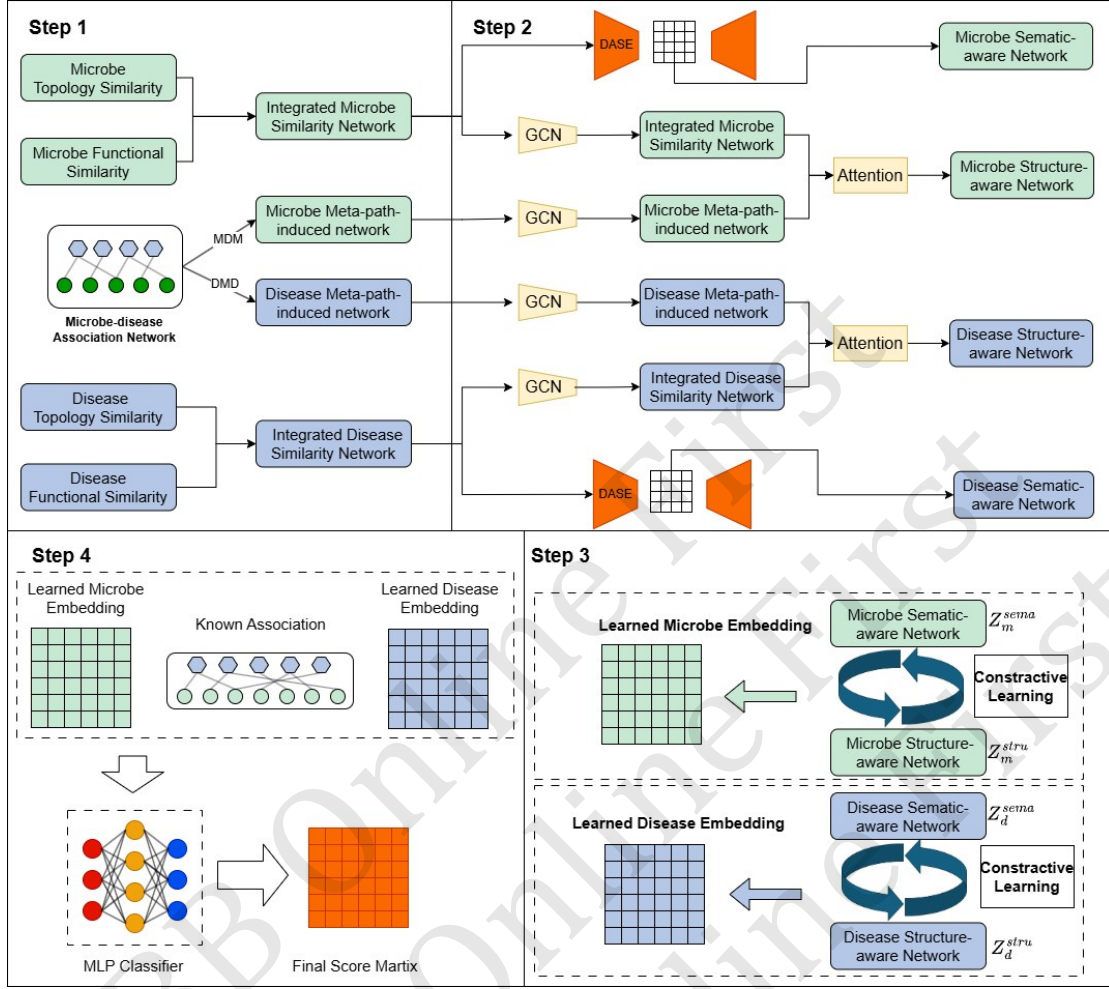


Fig. 2 The overall workflow of GECLMDA.

In step 1, microbe and disease similarity network are adopted, and microbe-disease association network is used to construct first-order meta-paths. In step 2, structural-aware representations are learned *via* node-level GCN and meta-path-based GCN and fused by an attention mechanism, while semantic-aware representations are obtained using a deep sparse autoencoder. In step 3, contrastive learning is applied to align topological and functional views for microbes and diseases. In step 4, the learned embeddings are finally used to predict potential microbe-disease associations.

at l^{th} , $H^{(l)}$ denotes the representations of nodes at l^{th} layer. The output representations of nodes at l^{th} layer can be input into the next layer of GCNs. In this way, we can get the nodes embeddings at any layer. Applying this operation to the microbe and disease similarity networks (M_s, D_s) yields node-level structural-aware representations GCN_m^{sn} and GCN_d^{sn} , respectively.

1.3.1.2 Meta-path-based graph convolution

To capture higher-order interaction semantics, we further construct meta-path-induced network based on the microbe-disease association network, namely $A_{MP^m}(M-D-M)$ and $A_{MP^d}(D-M-D)$. We apply GCNs on meta-path to learn the embeddings of microbes and diseases respectively, which can be calculated as:

$$H_{A_{MP^t}}^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A}_{MP^t} \tilde{D}^{-\frac{1}{2}} H_{A_{MP^t}}^{(l)} W_{A_{MP^t}}^{(l)} \right), t \in \{m, d\} \quad (4)$$

where $H_{A_{MP^t}}^{(l)}$ denotes the node representations learned from the corresponding meta-path network. The resulting embeddings are denoted as GCN_m^{mp} and GCN_d^{mp} .

1.3.1.3 Attention-based structural fusion

We employ an attention mechanism to fuse the two representations to integrate node-level and meta-path structural information:

$$z_t^{stru} = \alpha_t \cdot GCN_t^{sn} + (1 - \alpha_t) \cdot GCN_t^{mp}, t \in (m, d) \quad (5)$$

where α_t is a learnable attention weight. The fused embeddings z_m^{stru} and z_d^{stru} represent the final structural representations for microbes and diseases.

1.3.2 Semantic-aware representation learning via deep sparse autoencoder

The deep sparse autoencoder (DSAE) is an unsupervised neural network that updates parameters iteratively by minimizing reconstruction error between output and input, enabling efficient data compression and discriminative feature extraction^[35]. We adopt DSAE to learn semantic representations directly from the integrated similarity networks. Given the input similarity features, the encoder maps them into a low-dimensional sparse representation, which is reconstructed by the decoder. The loss function of the DSAE is defined as:

$$L_E = \frac{1}{N} \sum_{i=1}^N \|x_i - g(f(x_i))\|_2^2 + \beta \sum_{j=1}^S KL(\rho | \hat{\rho}_j) \quad (6)$$

where x_i denotes the input feature vector of the i -th node, $f(\cdot)$ and $g(\cdot)$ represent the encoder and decoder, respectively, ρ is the target sparsity level, is the average activation of the j -th hidden unit, β controls the sparsity penalty, and $KL(\cdot // \cdot)$ denotes the Kullback - Leibler divergence. The final semantic-aware embeddings for microbes and diseases are denoted as z_m^{sema} and z_d^{sema} , respectively.

1.3.3 Contrastive learning for feature refinement

Contrastive learning is a self-supervised representation learning paradigm that encourages similar samples to be mapped closer in the embedding space while pushing dissimilar samples apart^[35]. By maximizing agreement between positive pairs and minimizing similarity with negative samples, contrastive learning has shown strong effectiveness in similarity-driven tasks such as classification and clustering. In this work, we employ contrastive learning to bridge the gap between structural-aware and semantic-aware representations of microbes and diseases, thereby enhancing the consistency and discriminability of the learned embeddings.

Unlike conventional contrastive learning approaches that construct positive samples based on semantic or structural consistency within a single branch, our framework does not assume interaction knowledge inside each individual view. Instead, we adopt a cross-view positive sample selection strategy guided by meta-path-induced structural relationships^[29]. Specifically, for each microbe node m_i , its representations learned from the structural-aware branch and the semantic-aware branch naturally form an innate positive pair, i.e., $(z_{m_i}^{stru}, z_{m_i}^{sema})$.

Furthermore, to enrich the positive sample set, we exploit the meta-path-induced network to identify microbes that are indirectly associated with m_i . These candidate microbes are ranked according to the meta-

path connectivity scores in the MDM matrix, and the top- k (K_{pos}) nodes are selected as conditional positive samples, forming the positive set P_{m_i} . All remaining microbe nodes excluding P_{m_i} are treated as negative samples, constituting the set. The same strategy is applied symmetrically to disease nodes using the corresponding DMD matrix.

$$L_m = -\frac{1}{N_m} \sum_{i=1}^{N_m} \log \frac{\sum_{j \in P_{m_i}} \exp\left(\frac{\text{sim}(z_{m_i}^{stru}, z_{m_i}^{sema})}{\tau}\right)}{\sum_{k=1}^{N_m} \exp\left(\frac{\text{sim}(z_{m_i}^{stru}, z_{m_i}^{sema})}{\tau}\right)}$$

$$L_d = -\frac{1}{N_d} \sum_{i=1}^{N_d} \log \frac{\sum_{j \in P_{d_i}} \exp\left(\frac{\text{sim}(z_{d_i}^{stru}, z_{d_i}^{sema})}{\tau}\right)}{\sum_{k=1}^{N_d} \exp\left(\frac{\text{sim}(z_{d_i}^{stru}, z_{d_i}^{sema})}{\tau}\right)} \quad (7)$$

where $\text{sim}(z_{t_i}^{stru}, z_{t_j}^{sema})$, $t \in \{m, d\}$ is the cosine similarity between structural-aware network embedding and semantic-aware network embedding, and m and d means microbe and disease, τ is the temperature parameter for contrastive alignment, and N_t , $t \in \{m, d\}$ represents the total number of type t nodes. The total loss of contrast learning L_C can be represented as:

$$L_C = \lambda L_m + (1 - \lambda) L_d \quad (8)$$

where parameter λ is the coefficient to balance the contributions of the similarity network and meta-path-induced network.

1.3.4 Prediction and optimization

In this study, MLP is adopted as the final decoder. Firstly, the optimized microbe and disease embeddings obtained from the proposed framework are used as inputs to the MLP to predict potential microbe-disease associations. Finally, the association probability score $\hat{y}_{ij}$ between microbe m_i and disease d_j can be represented as:

$$\hat{y}_{ij} = \text{Sigmoid}(W_1(\text{ReLU}(W_2(z_{m_i} \odot z_{d_j})))) \quad (9)$$

where z_{m_i} and z_{d_j} means the embeddings of microbe m_i and disease d_j . The operation $\odot$ denotes

the element-wise multiplication between microbe embedding $z_{m_i} \in R^{F'}$ and disease embedding $z_{d_j} \in R^{F'}$, and $W_1 \in R^{1 \times F'}$, $W_2 \in R^{F' \times F'}$ are the learnable matrices.

We adopt the binary cross-entropy loss to train the model for microbe-disease association prediction. The binary cross-entropy loss used in GECLMDA is defined as:

$$L_B = - \sum_{(ij) \in S^+ \cup S^-} W_d y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \quad (10)$$

$$W_d = N_N / N_P \quad (11)$$

where (ij) denotes the microbe-disease sample for microbe m_i and disease d_j , S^+ and S^- are positive and

negative microbe-disease sample used to train, respectively. W_d is a weight to reduce the sample sparsity of dataset, N_N are the number negative samples, and N_p are positive. The overall training

objective of the proposed model is formulated as:

$$L = L_c + L_B + L_E \quad (12)$$

The detailed steps of GECLMDA are summarized in Table 2.

Table 2 The algorithm for microbe – disease association prediction

GECLMDA algorithm	
Input: initial interaction matrix A	
Output: final predict matrix S	
1:	Compute the disease similarity matrix D_d and the microbe similarity matrix M_s
2:	Compute the adjacency matrix of the disease graph $\widehat{D}_d$ and microbe graph $\widehat{M}_s$, respectively
3:	repeat
4:	Learn latent representations from the structural-aware and semantic-aware using graph convolutional networks and a deep sparse autoencoder, respectively (expression here is defined in equation (5) and (6))
5:	Construct positive and negative sample pairs and compute the contrastive learning loss based on reconstructed embeddings (expression here is defined in equation (7))
6:	Optimize the model parameters by minimizing the contrastive learning objective using the Adam optimizer
7:	until convergence or maximum epochs reached
8:	Predict potential microbe – disease associations using an MLP decoder based on the final learned embeddings
9:	return final predict matrix S

2 Result

2.1 Experimental settings and evaluation metrics

To evaluate the performance of GECLMDA and baseline methods, we employ 5-fold cross-validation (5-CV) on the HMDAD and Disbiome datasets. We further conducted three validation strategies to simulate distinct prediction scenarios: (1) a global test, where random associations are masked to assess overall predictive capacity; (2) a horizontal test, in which entire rows (diseases) are masked to evaluate the model's ability to predict associations for novel diseases; and (3) a vertical test, where entire columns (microbes) are masked to simulate the prediction of novel microbes. Each experiment is repeated 10 times to mitigate sampling variability.

2.1.1 Parameter settings

Key hyperparameters of GECLMDA, including the number of GCN and MLP layers (n_g and n_p), embedding dimension (k), sparsity coefficient (β), and number of training iterations (n_t), are optimized *via* 5-fold cross-validation under the global test setting on the HMDAD dataset under grid search. As shown in Fig. 3, we test GCN/MLP layers (1 – 3) and embedding sizes (32 – 512). Optimal performance is achieved with 1 GCN layer ($n_g=1$), 3 MLP layers ($n_p=3$), an embedding dimension of $k=128$, $\beta=0.1$, $n_t=1000$ training iterations, $k_{pos}=10$ and $\tau=0.8$.

The experiments were conducted in a Python 3.7.12 environment using PyTorch 1.12.1 with

CUDA, scikit-learn 0.22.1, and NumPy 1.21.6. All analyses are performed on a Windows 10 system equipped with an Intel(R) Core (TM) i5-8600KF CPU and an NVIDIA GeForce GTX 1060 6GB GPU.

2.1.2 Performance evaluation

To comprehensively evaluate the effectiveness of the proposed GECLMDA, we compared it with eight representative baseline methods covering network-based models (KATZHMMA^[12]), traditional machine learning approaches (LRLSHMDA^[36]), and deep learning-based frameworks (NTSHMDA^[37], DMCFMDA^[10], SARMDA^[38], GCDSAEMDA^[20], MSigVGAE^[39], and DVAMDA^[21]).

All experiments were performed on two widely used benchmark datasets, HMDAD and Disbiome, under three evaluation settings: global, disease-wise (horizontal), and microbe-wise (vertical) 5-fold cross-validation. For global prediction, the full network topology is retained. For unseen-disease and unseen-microbe evaluation, connections involving the held-out entities are removed when constructing the network topology. Model performance was assessed using five metrics^[40-41]: AUC, AUPR, Accuracy (ACC) Sensitivity, and Specificity, which jointly reflect ranking capability, robustness to class imbalance, and classification performance from a biological perspective.

2.1.2.1 Performance under global cross-validation

Under the global cross-validation setting, all known microbe-disease associations were randomly

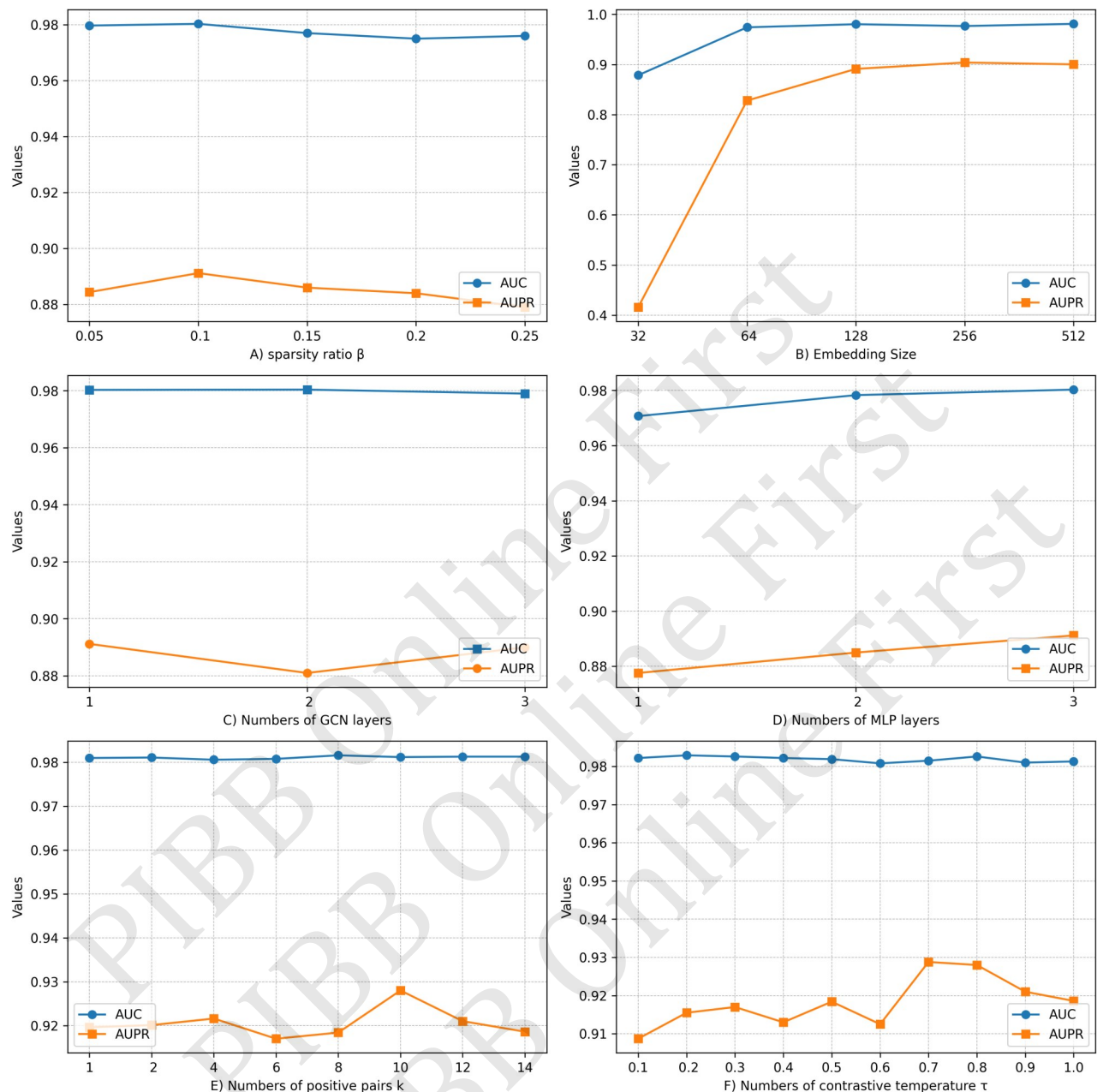


Fig. 3 Parameter sensitivity under 5-fold CV on the HMDAD dataset

partitioned into five folds, with four folds used for training and the remaining fold for testing in each round. This setting evaluates the overall predictive performance when both microbes and diseases in the test set have been observed during training.

As shown in Figure 4 and Table S1, GECLMDA achieves competitive performance against most existing state-of-the-art methods under the global cross-validation setting. On the HMDAD dataset, GECLMDA achieves an AUC of 0.9803 ± 0.0069 and an $AUPR$ of 0.9280 ± 0.0117 , demonstrating

strong ranking capability in prioritizing true associations. Although some baseline methods achieve higher scores on specific metrics, GECLMDA maintains a favorable balance between sensitivity and specificity, reflecting its robustness under highly imbalanced association data. On the Disbiome dataset, GECLMDA also performs competitively, achieving an AUC of 0.9446 ± 0.0083 and a high specificity of 0.9946 ± 0.0049 . The consistently high specificity across both datasets indicates that GECLMDA effectively suppresses false-positive predictions,

which is particularly important in microbe – disease association studies where experimental validation is costly. Overall, these results demonstrate that

GECLMDA can reliably integrate heterogeneous information to produce accurate and stable predictions in global settings.

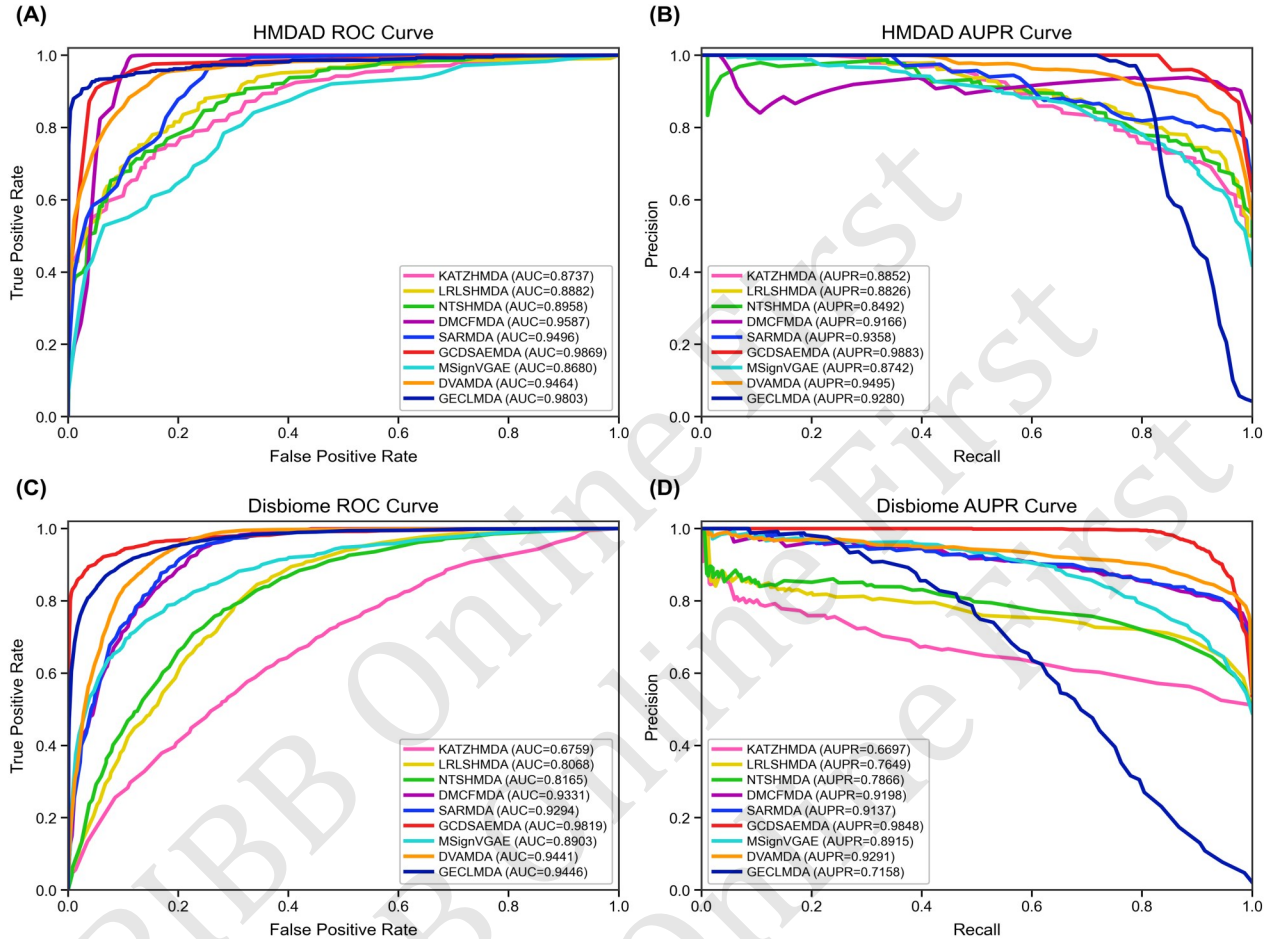


Fig. 4 The ROC and AUPR curves of GECLMDA as well as the baseline methods for predicting microbe–disease associations on HMDAD and Disbiome

2.1.2.2 Performance under disease-wise cross-validation

Disease-wise cross-validation evaluates the model's generalization ability to unseen diseases, representing a challenging disease cold-start setting. In this setting, 80% of diseases are used for training, while the remaining 20% are reserved for testing, with all associations related to test diseases removed from the training set.

GECLMDA maintains competitive ranking performance as shown in Table S2. On the HMDAD dataset, GECLMDA achieves an AUC of $0.899\ 0 \pm 0.003\ 4$ and an $AUPR$ of $0.931\ 3 \pm 0.004\ 0$, indicating that the model can effectively prioritize relevant microbes for diseases without prior association information. While sensitivity is relatively lower in this setting, GECLMDA exhibits high specificity,

suggesting a conservative prediction behavior that reduces false discoveries under disease-level cold-start conditions. On the Disbiome dataset, disease-wise prediction is particularly difficult due to extreme sparsity and heterogeneous disease definitions. Despite this, GECLMDA attains a remarkably high AUC of $0.976\ 4 \pm 0.002\ 8$, demonstrating strong discriminative ability at the ranking level. However, the lower AUPR and sensitivity suggest a reduced ability to identify positive associations. This may partly reflect the limited relational information available for unseen diseases after disease-wise partitioning, posing greater challenges for positive-association prediction.

2.1.2.3 Performance under microbe-wise cross-validation

Microbe-wise cross-validation further assesses

the model's ability to predict associations for previously unseen microbes, which is particularly relevant for discovering disease relevance of newly identified or poorly characterized microbial species. In this setting, 80% of microbes are used for training and the remaining 20% for testing.

As shown in Table S3, GECLMDA demonstrates stable performance under microbe-wise evaluation. On the HMDAD dataset, GECLMDA achieves an AUC of 0.9816 ± 0.0153 and an $AUPR$ of 0.9790 ± 0.0064 , indicating strong generalization capability even when no prior association information is available for test microbes. The balanced performance between sensitivity and specificity suggests that the model can effectively identify potential associations while maintaining control over false positives.

Similarly, on the Disbiome dataset, GECLMDA achieves an AUC of 0.9466 ± 0.0091 and maintains high specificity, while the $AUPR$ (0.6316) and sensitivity (0.5323) suggest that positive associations remain difficult to identify in the microbe-wise cold-start setting.

2.1.2.4 Validation on the Peryton dataset

Furthermore, we conducted additional global five-fold cross-validation experiments on the Peryton dataset. As shown in Figure 5, GECLMDA achieved an AUC of 0.9546 ± 0.0067 and an $AUPR$ of 0.9583 ± 0.0063 , only slightly lower than GCDSAEMDA. Nevertheless, GECLMDA maintained consistently strong ranking performance on Peryton, providing additional evidence of its applicability across different microbe-disease association datasets.

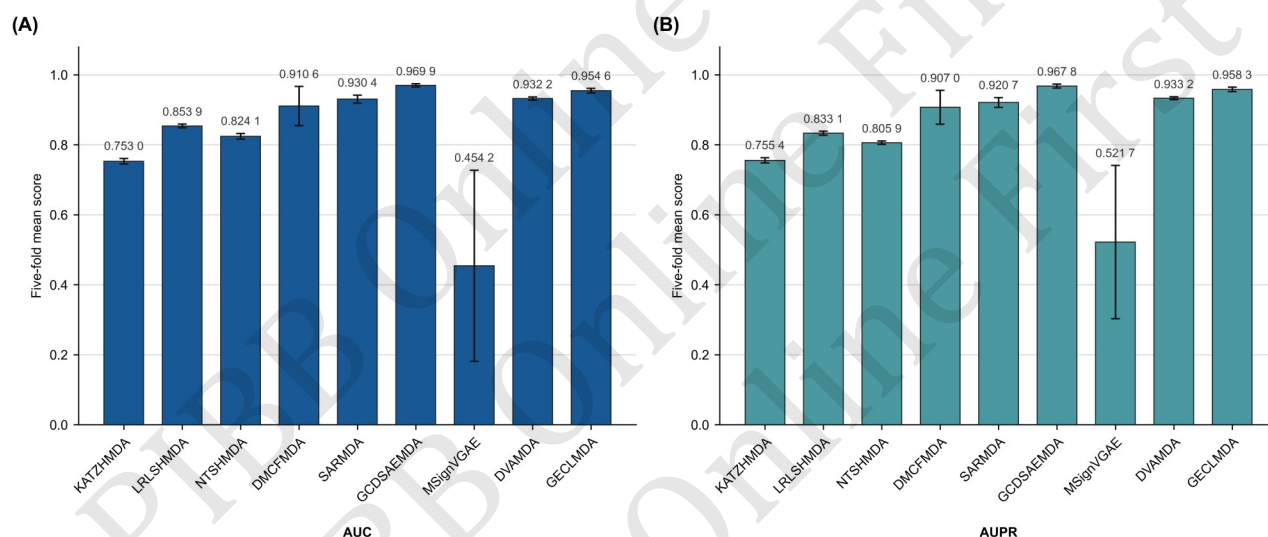


Fig. 5 Performance comparison of different methods on the Peryton dataset under five-fold cross-validation

2.2 Ablation experiments

To investigate the contribution of individual components in GECLMDA, we conduct a series of ablation experiments on the HMDAD and Disbiome datasets. Specifically, these experiments are designed to evaluate (1) the necessity of jointly modeling structural and semantic information, (2) the effectiveness of the contrastive alignment mechanism, and (3) the robustness of the overall framework under different evaluation settings. All ablation studies are performed under the same experimental settings to ensure fair comparison.

2.2.1 Performance comparison with GCN and deep sparse autoencoder

To assess the individual contributions of the structural-aware and semantic-aware branches, we

conduct ablation experiments on the HMDAD and Disbiome datasets without contrastive learning. Specifically, we compare GECLMDA with two simplified variants: an Autoencoder variant that retains only the semantic-aware representation learning branch, and a GCN variant that retains only the semantic representation learning branch. The GCN variant uses the same one-layer GCN setting as GECLMDA and includes the meta-path-induced networks. Three comparative strategies are designed to evaluate the individual impact of each component on model performance.

As shown in Table 3, our method demonstrates strong performance, effectively capturing key features and structural information from structural-aware embeddings. Notably, the deep sparse autoencoder

exhibits robust predictive capability, while the performance of GCNs appears to be highly sensitive

to the data partitioning strategy.

Table 3 Performance of different branches under the global test of 5-fold CV on HMDAD and Disbiome

Datasets	Methods	ACC	AUC	AUPR
HMDAD	Autoencoder	0.986 0±0.001 1	0.983 7±0.003 2	0.893 7±0.012 8
	GCN	0.985 6±0.001 7	0.975 0±0.008 9	0.881 0±0.013 5
	GECLMDA	0.989 1±0.001 5	0.980 3±0.006 9	0.928 0±0.011 7
Disbiome	Autoencoder	0.956 0±0.002 1	0.943 2±0.021 2	0.693 2±0.012 7
	GCN	0.933 1±0.031 7	0.925 3±0.024 9	0.641 0±0.016 5
	GECLMDA	0.987 1±0.001 5	0.944 6±0.008 3	0.715 8±0.001 2

The best and second-best results are written in bold and underlined, respectively.

2.2.2 Performance comparison with different representation components

To investigate the contribution of each representation component, we perform component-wise ablation experiments by removing DSAE, sn-GCN, or mp-GCN from the full model:

w/o DSAE: removes the semantic-aware representation component and retains the two graph-based structural components.

w/o sn-GCN: removes the similarity-network graph convolution component while retaining mp-GCN and DSAE.

w/o mp-GCN: removes the meta-path-based graph convolution component while retaining sn-GCN and DSAE.

As shown in Table 4, the contribution of each component varies across datasets and metrics. GECLMDA achieves the best overall performance on HMDAD, while on Disbiome the full model obtains the highest AUPR despite slightly lower ACC and AUC than the w/o mp-GCN variant. These results indicate that the three components provide complementary structural and semantic information rather than uniform gains across all metrics.

Table 4 Performance of different components under the global test of 5-fold CV on HMDAD and Disbiome

Datasets	Methods	ACC	AUC	AUPR
HMDAD	w/o DSAE	0.988 6±0.002 8	0.976 6±0.008 0	0.908 5±0.016 8
	w/o sn-GCN	0.987 9±0.002 1	0.969 5±0.016 5	0.847 8±0.024 0
	w/o mp-GCN	0.987 6±0.003 6	0.971 8±0.012 0	0.911 2±0.016 0
	GECLMDA	0.989 1±0.001 5	0.980 3±0.006 9	0.928 0±0.011 7
Disbiome	w/o DSAE	0.896 8±0.001 3	0.920 9±0.013 2	0.359 7±0.218 4
	w/o sn-GCN	0.971 0±0.004 9	0.950 5±0.006 2	0.399 6±0.025 3
	w/o mp-GCN	0.987 6±0.001 8	0.947 1±0.029 1	0.700 1±0.037 6
	GECLMDA	0.987 1±0.001 5	0.944 6±0.008 3	0.715 8±0.001 2

The best and second-best results are written in bold and underlined, respectively.

2.2.3 Performance comparison with different contrastive learning strategies

To further evaluate the role of contrastive learning, we design three comparative strategies that differ in how cross-view consistency is enforced:

Strategy 1 (full model): positive pairs are defined based on cross-view counterparts and meta-path - induced conditional similarities, enabling structural-aware alignment between representations from different views.

Strategy 2 (w/o CL): completely removes the contrastive learning objective and directly concatenates the representations produced by the two

encoders for downstream prediction

Strategy 3 (CL-Random): replaces the structured positive sample selection with random sampling, constructing positive pairs without considering structural or semantic relevance.

As shown in Fig. 6, the results demonstrate that removing contrastive learning leads to a consistent decline in performance, underscoring its importance in effectively aligning heterogeneous representations. Although CL-Random yields slight improvements over the non-contrastive baseline, its performance remains inferior to that of the proposed strategy. This observation indicates that contrastive learning alone is

insufficient; instead, carefully designed positive sample selection grounded in network structure is critical for effective cross-view alignment. Overall,

these findings confirm that the proposed contrastive learning strategy plays a central role in enhancing representation.

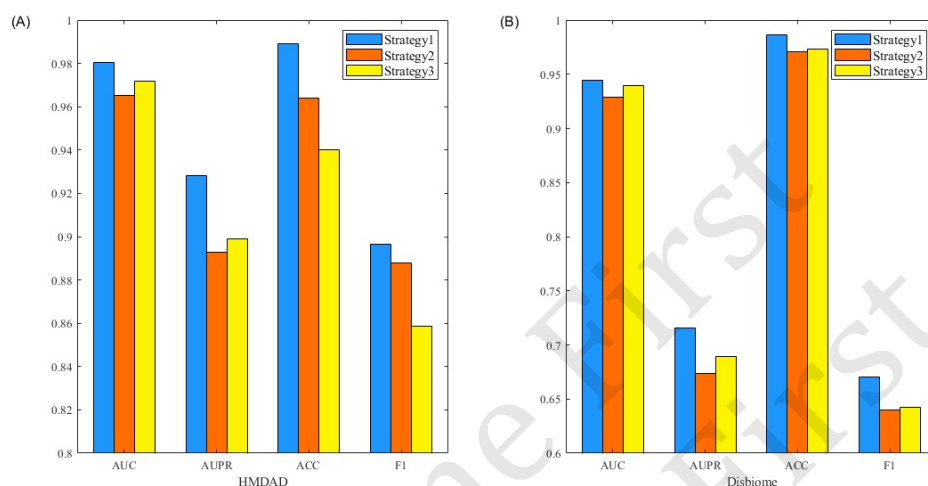


Fig. 6 the performance of three different strategies on HMDAD and Disbiome

2.3 Case study

To further assess the predictive capability of GECLMDA, we trained the model on the HMDAD dataset and applied it to infer association scores between specific microbes and diseases. Based on the predicted probabilities, we identify the top 15 candidate diseases associated with two conditions: asthma and inflammatory bowel disease (IBD). These predictions were subsequently validated through literature searches on PubMed.

Asthma is a common chronic inflammatory airway disease affecting over 300 million individuals globally, characterized by heterogeneous clinical manifestations such as recurrent wheezing, chest tightness, dyspnea, and coughing^[42]. Numerous studies have highlighted the involvement of genetic, environmental, and microbial factors—particularly from the gut and respiratory tract—in asthma pathogenesis^[43]. For example, Vael *et al.*^[44] found that *Clostridium* is significantly associated with asthma. Fujimura *et al.*^[45] found that reduced abundance of *Akkermansia*, *Bifidobacterium*, and *Faecalibacterium prausnitzii* in the gut microbiota of children is associated with allergic asthma. As shown in Table 5, GECLMDA predicted the top 15 microbes potentially associated with asthma, 10 of which have been supported by previous studies.

IBD refers to a group of chronic, relapsing inflammatory disorders of the gastrointestinal tract,

primarily including Crohn's disease, ulcerative colitis, and indeterminate colitis^[46]. IBD imposes a substantial disease burden, often severely impairing quality of life. Its etiology is widely considered to involve complex interactions among the microbiome, host immunity, and environmental factors. Several microbial alterations have been implicated in IBD, including reduced *Bacteroidetes* abundance, increased *Ruminococcus gnavus*, and disrupted host – microbe homeostasis^[47-49]. As presented in Table 6, GECLMDA successfully identified 15 IBD-associated candidate microbes, many of which are supported by existing literature, further demonstrating the model's effectiveness in uncovering disease-related microbial signatures.

3 Discussion

In this study, we developed a graph-enhanced contrastive learning framework for microbe – disease association prediction. It effectively integrates complementary structural and semantic information for microbe – disease association prediction. By disentangling structural-aware learning from semantic-aware learning and aligning them through contrastive optimization, GECLMDA enables the learned representations to simultaneously preserve interaction patterns and semantic characteristics. Different from methods that directly contrast multiple graph-derived representations, GECLMDA first consolidates graph-

Table 5 The predicted top 15 microbes associated with Asthma

Rank	Microbe	Evidence
1	Sphingomonadaceae	PMID: 21194740
2	Helicobacter pylori	PMID: 28389738
3	Comamonadaceae	PMID: 21194740
4	Oxalobacteraceae	PMID: 21194740
5	Bacteroidetes	PMID: 25306600
6	Proteobacteria	PMID: 35685081
7	Prevotella	PMID: 34612663
8	Actinobacteria	PMID: 23265859
9	Clostridia	Unconfirmed
10	Haemophilus	PMID: 35904980
11	Shigella dysenteriae	Unconfirmed
12	Actinomyces	Unconfirmed
13	Veillonella	PMID: 26424567
14	Odoribacter	Unconfirmed
15	Alistipes finegoldii	Unconfirmed

Table 6 The predicted top 15 microbes associated with IBD

Rank	Microbe	Evidence
1	Haemophilus	PMID: 33666710
2	Lachnospiraceae	PMID: 17699621
3	Bacteroides vulgatus	PMID: 29454108
4	Streptococcus	PMID: 30392911
5	Veillonella	PMID: 16868828
6	Actinobacteria	PMID: 17699621
7	Verrucomicrobiaceae	PMID: 22572638
8	Prevotella	Unconfirmed
9	Escherichia coli	PMID: 19573336
10	Bacteroides	Unconfirmed
11	Clostridium coccoides	PMID: 27687331
12	Proteobacteria	Unconfirmed
13	Lactobacillus	Unconfirmed
14	Bifidobacterium	PMID: 34337079
15	Faecalibacterium prausnitzii	PMID: 25391346

based relational information before performing cross-representation alignment, thereby enabling the joint exploitation of relational structures and latent similarity patterns. This design alleviates the limitations of existing methods that rely on a single information source or loosely coupled multi-view fusion strategies.

Extensive experiments on two benchmark datasets under global, disease-wise, and microbe-wise cross-validation settings demonstrate that GECLMDA consistently achieves strong performance across

multiple evaluation metrics, including *AUC*, *AUPR*, accuracy, sensitivity, and specificity. Additional global validation on Peryton further supports the generalizability of GECLMDA. In particular, its robust performance in disease-wise and microbe-wise evaluations highlights its capability of the generalization, which is critical for discovering associations involving previously uncharacterized diseases or microbial species. Case studies further indicate that GECLMDA can generate biologically plausible and potentially novel predictions, underscoring its practical value in guiding downstream experimental validation.

Despite these results, several limitations remain. The sparsity and imbalance of known microbe - disease associations inevitably constrain model training and may bias predictions toward well-studied entities. Although the contrastive learning strategy partially mitigates this issue by enhancing representation robustness, incorporating more advanced self-supervised or imbalance-aware learning mechanisms may further improve performance. However, prediction performance still varies across datasets and evaluation scenarios, particularly in challenging cold-start settings. In addition, the current framework primarily relies on similarity-derived networks and does not explicitly integrate rich biological priors, such as microbial abundance profiles, metabolic pathways, genomic features, or disease-related drugs and metabolites. Future work will extend GECLMDA by incorporating heterogeneous biological data and more expressive graph modeling strategies to improve predictive performance, interpretability, and biological relevance. Further studies will also investigate more robust strategies for class imbalance, potential false-negative associations, and threshold selection in different prediction scenarios.

4 Conclusion

The gut microbiome is increasingly recognized as a potential indicator of disease, highlighting the importance of characterizing disease-microbe relationships. Integrating complementary biological information can improve association prediction. Our framework further supports this hypothesis by combining graph topology and semantic information, offering a promising computational approach for identifying disease-associated microbes.

Data Availability Statement The source and datasets analyzed during the current study are

available at <https://github.com/GUT-ShiLab/GECLMAD>. In addition, the associated data for this paper (DOI: 10.57760/sciencedb.015pj / CSTR: 31253.11. sciencedb.015pj) can be accessed and obtained from the Science Data Bank database (<https://www.scidb.cn/s/zqIvim>).

Supplementary Available online (<http://www.pibb.ac.cn>, <http://www.cnki.net>):

PIBB_20260325_Table_S1.pdf

PIBB_20260325_Table_S2.pdf

PIBB_20260325_Table_S3.pdf

References

- [1] Joos R, Boucher K, Lavelle A, *et al.* Examining the healthy human microbiome concept. *Nat Rev Microbiol*, 2025, **23**(3): 192-205
- [2] Zhou X, Shen X, Johnson J S, *et al.* Longitudinal profiling of the microbiome at four body sites reveals core stability and individualized dynamics during health and disease. *Cell Host Microbe*, 2024, **32**(4): 506-526.e9
- [3] Ross F C, Patangia D, Grimaud G, *et al.* The interplay between diet and the gut microbiome: implications for health and disease. *Nat Rev Microbiol*, 2024, **22**(11): 671-686
- [4] Potiron A, Francken J C, El Aidy S. Rethinking microbiome health through functional dynamics. *Cell Host Microbe*, 2026, **34**(4): 562-566
- [5] Corral López R, Bonachela J A, Dominguez-Bello M G, *et al.* Imbalance in gut microbial interactions as a marker of health and disease. *Science*, 2026, **391**(6788): 890-895
- [6] Metwaly A, Kriaa A, Hassani Z, *et al.* A Consensus Statement on establishing causality, therapeutic applications and the use of preclinical models in microbiome research. *Nat Rev Gastroenterol Hepatol*, 2025, **22**(5): 343-356
- [7] Gilbert J A, Azad M B, Bäckhed F, *et al.* Clinical translation of microbiome research. *Nat Med*, 2025, **31**(4): 1099-1113
- [8] Chen H, Chen K. Ensemble learning based on matrix completion improves microbe-disease association prediction. *Brief Bioinform*, 2025, **26**(2): bbaf075
- [9] Shi K, Huang K, Li L, *et al.* Predicting microbe-disease association based on graph autoencoder and inductive matrix completion with multi-similarities fusion. *Front Microbiol*, 2024, **15**: 1438942
- [10] Wu Y, Hu M, Zhu J. DMCFMDA: a dual-channel multi-source cross-modal fusion network with contrastive learning for microbe - disease prediction. *Biomed Signal Process Control*, 2025, **110**: 108039
- [11] Mishra S K, Praba J J, Gurung K, *et al.* A multipurpose machine learning application in microbiological data. *CBIO*, 2026, **21**(6): 445-463
- [12] Chen X, Huang Y A, You Z H, *et al.* A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics*, 2017, **33**(5): 733-739
- [13] Liu Y, Wang S L, Zhang J F, *et al.* DMFMDA: prediction of microbe-disease associations based on deep matrix factorization using Bayesian personalized ranking. *IEEE*, 2021, **18**(5): 1763-1772
- [14] Wang L, Wang Y, Li H, *et al.* A bidirectional label propagation based computational model for potential microbe-disease association prediction. *Front Microbiol*, 2019, **10**: 684
- [15] Liang C, Wang D, Zhang H, *et al.* Robust tensor subspace learning for incomplete multi-view clustering. *IEEE Trans Knowl Data Eng*, 2024, **36**(11): 6934-6948
- [16] Ma Y, Liu G, Ma Y, *et al.* Integrative analysis for identifying co-modules of microbe-disease data by matrix tri-factorization with phylogenetic information. *Front Genet*, 2020, **11**: 83
- [17] Peng L, Shen L, Liao L, *et al.* RNMFMDA: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front Microbiol*, 2020, **11**: 592430
- [18] Liu X W, Li H L, Ma C Y, *et al.* Predicting the role of the human gut microbiome in type 1 diabetes using machine-learning methods. *Brief Funct Genom*, 2024, **23**(4): 464-474
- [19] Peng L H, Yin J, Zhou L, *et al.* Human microbe-disease association prediction based on adaptive boosting. *Front Microbiol*, 2018, **9**: 2440
- [20] Wang B, Zhao W, Du X, *et al.* Predicting potential microbe-disease associations based on heterogeneous graph attention network and deep sparse autoencoder. *Eng Appl Artif Intell*, 2025, **147**: 110301
- [21] Ye Q, Sun Y. Harnessing dual variational autoencoders to decode microbe roles in diseases for traditional medicine discovery. *Front Pharmacol*, 2025, **16**: 1578140
- [22] Xing J, Shi Y, Su X, *et al.* Discovering microbe-disease associations with weighted graph convolution networks and taxonomy common tree. *CBIO*, 2024, **19**(7): 663-673
- [23] Shi K, Li L, Wang Z, *et al.* Identifying microbe-disease association based on graph convolutional attention network: Case study of liver cirrhosis and epilepsy. *Front Neurosci*, 2022, **16**: 1124315
- [24] Long Y, Luo J, Zhang Y, *et al.* Predicting human microbe-disease associations via graph attention networks with inductive matrix completion. *Brief Bioinform*, 2021, **22**(3): bbaf146
- [25] Qiao J, Gao W, Jin J, *et al.* Molecular pretraining models towards molecular property prediction. *Sci China Inf Sci*, 2025, **68**(7): 170104
- [26] Chen Y, Lei X. Metapath aggregated graph neural network and tripartite heterogeneous networks for microbe-disease prediction. *Front Microbiol*, 2022, **13**: 919380
- [27] Zhang L, Shang J, Kong X, *et al.* Two-stage multi-view graph spectral clustering for single-cell RNA-seq data. *Curr Genomics*, 2025, **26**(6): 673-689
- [28] Momanyi B M, Temesgen S A, Grace-Mercure B K, *et al.* RWRGDR: random walk and GraphSAGE-based framework for enhanced drug repositioning. *Curr Drug Targets*, 2026, **27**(7): 496-506
- [29] Tian Z, Yu Y, Fang H, *et al.* Predicting microbe - drug associations

- with structure-enhanced contrastive learning and self-paced negative sampling strategy. *Brief Bioinform*, 2023, **24**(2): bbac634
- [30] Lei S, Lei X, Chen M, *et al.* Drug repositioning based on deep sparse autoencoder and drug – disease similarity. *Interdiscip Sci Comput Life Sci*, 2024, **16**(1): 160-175
- [31] Li H, Wang Y, Jiang J, *et al.* A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front Microbiol*, 2019, **10**: 676
- [32] Janssens Y, Nielandt J, Bronselaer A, *et al.* Disbiome database: linking the microbiome to disease. *BMC Microbiol*, 2018, **18** (1): 50
- [33] Qi C, Cai Y, Qian K, *et al.* gutMDisorder v2.0: a comprehensive database for dysbiosis of gut microbiota in phenotypes and interventions. *Nucleic Acids Res*, 2023, **51**(D1): D717-D722
- [34] Skoufos G, Kardaras F S, Alexiou A, *et al.* Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res*, 2021, **49**(D1): D1328-D1333
- [35] Liu X, Song C, Huang F, *et al.* GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief Bioinform*, 2022, **23**: bbab457
- [36] Wang F, Huang Z A, Chen X, *et al.* LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction. *Sci Rep*, 2017, **7**(1): 7601
- [37] Luo J, Long Y. NTSMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 2020, **17**(4): 1341-1351
- [38] He L, Zou Q, Dai Q, *et al.* Adversarial regularized autoencoder graph neural network for microbe-disease associations prediction. *Brief Bioinform*, 2024, **25**(6): bbae584
- [39] Zhu H, Hao H, Yu L. Identification of microbe – disease signed associations via multi-scale variational graph autoencoder based on signed message propagation. *BMC Biol*, 2024, **22**(1): 172
- [40] Huang Z, Xiao Z, Ao C, *et al.* Computational approaches for predicting drug-disease associations: a comprehensive review. *Front Comput Sci*, 2024, **19**(5): 195909
- [41] Guo X, Huang Z, Ju F, *et al.* Highly accurate estimation of cell type abundance in bulk tissues based on single-cell reference and domain adaptive matching. *Adv Sci (Weinh)*, 2024, **11**(7): e2306329
- [42] Zhang Z Q, Li J Y, Fu C Y, *et al.* Global, regional and national burden of asthma attributable to metabolic diseases from 1990 to 2021 and projected trends to 2040. *BMC Pulm Med*, 2025, **25** (1): 546
- [43] Chou C K, Winker R, Rebuli M E, *et al.* Respiratory health impacts from natural disasters and other extreme weather events: the role of environmental stressors on asthma and allergies. *Curr Allergy Asthma Rep*, 2025, **25**(1): 25
- [44] Vael C, Vanheirstraeten L, Desager K N, *et al.* Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol*, 2011, **11**: 68
- [45] Fujimura K E, Sitarik A R, Havstad S, *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat Med*, 2016, **22**(10): 1187-1191
- [46] Han T, Zhang Y, Zheng G, *et al.* From pathogenic mechanisms to therapeutic perspectives: a review of gut microbiota and intestinal mucosal immunity in inflammatory bowel disease. *Front Immunol*, 2025, **16**: 1704651
- [47] Florio M, Crudele L, Sallustio F, *et al.* Disentangling the nutrition-microbiota liaison in inflammatory bowel disease. *Mol Aspects Med*, 2025, **102**: 101349
- [48] Iliev I D, Ananthakrishnan A N, Guo C J. Microbiota in inflammatory bowel disease: mechanisms of disease and therapeutic opportunities. *Nat Rev Microbiol*, 2025, **23**(8): 509-524
- [49] Bethlehem L, Estevinho M M, Grinspan A, *et al.* Microbiota therapeutics for inflammatory bowel disease: the way forward. *Lancet Gastroenterol Hepatol*, 2024, **9**(5): 476-486

基于结构语义图对比学习的微生物-疾病关联预测*

石凯^{1,2,3)} 许佑昂¹⁾ 黄凯¹⁾ 冀庆荣¹⁾ 曹瑞¹⁾ 汪彦君¹⁾ 蒋琼^{4)**}

(¹⁾ 桂林理工大学计算机科学与工程学院, 桂林 541004; (²⁾ 桂林理工大学广西嵌入式技术与智能系统重点实验室, 桂林 541004;

(³⁾ 广西人工智能学院, 南宁 530028; (⁴⁾ 桂林理工大学物理与电子信息工程学院, 桂林 541004)

摘要 **目的** 肠道微生物失衡与多种疾病的发生发展密切相关。现有微生物-疾病关联预测方法对多源异构生物信息利用不充分, 且已知关联的噪声与不完整性导致预测稳定性和泛化能力不足。**方法** 本文提出一种基于图增强对比学习的微生物-疾病关联预测框架。该框架通过融合结构学习和语义学习来获取高质量的节点表示。该框架利用图卷积网络捕获微生物与疾病之间的结构关系, 并采用深度稀疏自编码器从相似性信息中提取潜在语义特征。此外, 该框架还引入对比学习策略来对齐两种视图的表示, 并将融合后的嵌入用于预测潜在的微生物-疾病关联。在HMDAD、Disbiome和Peryton三个数据集上开展五折交叉验证, 通过消融实验和案例研究评价模型性能。**结果** 全局五折交叉验证结果表明, GECLMDA在HMDAD、Disbiome和Peryton数据集上的曲线下面积(AUC)分别为 0.9803 ± 0.0069 、 0.9446 ± 0.0083 和 0.9546 ± 0.0067 。在不同评价设置下, 该方法相较于多数现有方法表现出具有竞争力的预测性能; 消融实验表明, 结构学习、语义学习和对比学习均对模型预测性能具有积极贡献; 案例研究进一步表明, 该方法能够有效识别潜在的疾病相关微生物。**结论** 所提出的框架能有效提高微生物-疾病关联预测的稳定性和泛化能力, 可作为疾病相关微生物预测的潜在工具。

关键词 微生物-疾病关联, 关联预测, 元路径, 图卷积神经网络, 深度稀疏编码器, 对比学习

中图分类号 TP183, R37

DOI: 10.3724/j.pibb.2026.0325

CSTR: 32369.14.pibb.20260325

* 国家自然科学基金(62562022)和广西自然科学基金(2025JJA170175)资助项目。

** 通讯联系人。

石凯 Tel: 13737707647, E-mail: mail_shikai@foxmail.com

蒋琼 Tel: 0773-5896435, E-mail: huichao1228@163.com

收稿日期: 2026-06-15, 接受日期: 2026-10-17