

Article

DDHMDA: Dual Dynamic Hypergraph Convolution Framework for Human Microbe-Disease Association Prediction

Zhi Wu ^{1,2} and Zhaohui Liao ^{3,*} ¹ School of Big Data and Artificial Intelligence, Guangdong Polytechnic of Water Resources and Electric Engineering, Guangzhou 510925, China; wuzhi@gdsdxy.cn² Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia³ School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

* Correspondence: liaozhaohui@gdut.edu.cn

Abstract

The human microbiota is essential for maintaining physiological homeostasis, and microbial dysbiosis is increasingly implicated in the pathogenesis of complex diseases. Identifying potential microbe-disease associations (MDAs) can therefore facilitate mechanistic investigation, biomarker discovery, and therapeutic development. However, wet-laboratory validation is costly and time-consuming, while existing computational methods often struggle with sparse association networks and complex nonlinear interactions. We propose a novel deep learning approach named the Dual Dynamic Hypergraph Convolution Framework for Human Microbe-Disease Association Prediction (DDHMDA). Specifically, DDHMDA first utilizes graph convolutional networks to encode local topological features. Subsequently, it dynamically constructs a dual hypergraph architecture: a differentiable K-means similarity hypergraph to capture intra-modal global clustering patterns, and an attention-based cross-modal interaction hypergraph to model inter-modal interactions synergistically. Under leakage-free pair-level five-fold cross-validation (denoted as CV3), DDHMDA achieved AUC/AUPR values of $0.9789 \pm 0.0177/0.9843 \pm 0.0129$ on HMDAD and $0.9651 \pm 0.0042/0.9740 \pm 0.0031$ on Disbiome. DDHMDA also obtained the best overall CV3 performance among the eight evaluated methods. Furthermore, ablation experiments and case studies validate the practical effectiveness of individual modules and the biological interpretability in discovering novel MDAs. Therefore, DDHMDA would be a reliable tool for identifying potential MDAs.

Keywords: microbe; dynamic hypergraph convolution; microbe-disease association; association prediction



Academic Editor: Chun Chun Wang

Received: 31 March 2026

Revised: 22 June 2026

Accepted: 29 June 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

MSC: 68T07

1. Introduction

Microorganisms widely inhabit diverse ecological niches of the human body, including the gut, skin, oral cavity, and respiratory tract, forming a complex microbial ecosystem known as the human microbiome [1–4]. Advances in high-throughput sequencing and metagenomic analysis have revealed that microbial communities play essential roles in host metabolism, immune regulation, and maintenance of physiological homeostasis [2,5]. Microbes have emerged as key players in the pathogenesis of multiple human diseases [6].

Accumulating evidence demonstrates specific associations between microbes and human diseases [4,7–10]. For instance, *Fusobacterium nucleatum* promotes colorectal tumorigenesis through immune modulation and inflammatory signaling [11]. Altered gut microbiota profiles have been observed in Parkinson’s disease patients, suggesting microbiome involvement in neurodegeneration [12]. Moreover, it is revealed that enrichment of *Prevotella copri* has been associated with rheumatoid arthritis and host inflammatory activation [13]. Recent studies further indicate microbiome participation in metabolic and cardiovascular diseases through host-microbe metabolic interactions [14], and microbiota-derived metabolites have been implicated in modulating systemic immune responses [15]. In addition, microbiome perturbations have been linked to variability in cancer immunotherapy responses [16]. These findings collectively suggest that microbes participate in disease pathogenesis rather than acting as passive colonizers. Therefore, identifying potential microbe-disease associations (MDAs) is important for understanding disease mechanisms and improving diagnosis as well as therapeutic strategies. However, traditional biological experiments are costly and time-consuming, making it impractical to validate all candidate associations exhaustively. Computational prediction of MDAs provides an efficient complementary strategy by prioritizing candidate associations for experimental verification and enabling systematic exploration of MDAs.

Currently, researchers have proposed numerous computational model-based methods to predict the potential MDAs, and these previous models have achieved relatively ideal research results. However, there are still plenty of limitations. The performance of network propagation-based methods relies heavily on the completeness of the network structure, and their adaptability to sparse networks or scenarios involving new microbes and new diseases is limited. Matrix factorization and completion-based methods usually assume that the association relationships satisfy a linear low-rank structure, making it difficult to characterize the complex and nonlinear interactive relationships between microbes and diseases. Moreover, machine learning-based methods treat microbe-disease association prediction as a supervised or semi-supervised learning task, automatically modeling microbes, diseases, and their interactive features through feature engineering or representation learning. They possess stronger non-linear modeling capabilities and flexibility, but their performance often depends on high-quality feature construction and the scale of labeled samples, while their model complexity and training costs are relatively high.

To overcome the aforementioned limitations, particularly the difficulty in capturing high-order non-linear interactions and the vulnerability to sparse network structures, we propose a novel deep learning approach named the Dual Dynamic Hypergraph Convolution Framework for Human Microbe-Disease Association Prediction (DDHMDA). Unlike traditional graph-based methods that rely on fixed pairwise connections, DDHMDA dynamically learns high-order topological structures in an end-to-end framework. Specifically, the framework first encodes local features using a linear transformation and the GCN method. Subsequently, it constructs a dual dynamic hypergraph: a differentiable K-means similarity hypergraph to capture intra-modal global clustering patterns, and a cross-modal interaction hypergraph to model the synergistic effects between microbes and diseases. Finally, the model concatenates the four sets of embeddings for a given microbe-disease pair, and the fused feature embedding is fed into a prediction module to obtain the predicted association score. Extensive experiments evaluated via 5-fold cross-validation (CV) under three distinct settings (disease-level, denoted as CV1; microbe-level, denoted as CV2; and pair-level, denoted as CV3) demonstrate the robust generalization capability of DDHMDA across varying degrees of network sparsity and novel entity scenarios, namely HMDAD and Disbiome benchmark datasets. Herein, CV3 is used as the main evaluation setting, while the more stringent CV1 and CV2 cold-start results are reported in Appendix A.

Under CV3, DDHMDA achieves Area Under the ROC Curve (AUC)/Area Under the Precision-Recall Curve (AUPR) values of 0.9789/0.9843 on HMDAD and 0.9651/0.9740 on Disbiome, respectively. Ablation experiments quantify the contribution of each component, and the existing asthma and Inflammatory Bowel Disease (IBD) case studies provide complementary biological validation. The top-ranked candidate associations prioritized by DDHMDA were strongly corroborated by recent independent biological and clinical literature, demonstrating the model's high reliability and strong biological interpretability for real-world biomarker discovery.

The main contributions of DDHMDA are summarized as follows:

- Leakage-free evaluations on HMDAD and Disbiome show that DDHMDA achieves the best overall CV3 performance among the eight evaluated methods. The supplementary CV1 and CV2 experiments further characterize its behavior under disease-level and microbe-level cold-start conditions.
- We propose a dual dynamic hypergraph convolution network that jointly captures intra-modal high-order similarities and inter-modal interactions. Differentiable K-means clustering enables the similarity-hypergraph topology to be optimized during training instead of relying on a fixed, manually specified structure.
- We specifically design an interaction hypergraph mechanism where nodes from one modality attend to the hyperedge cluster centers of the other. This effectively captures the complex and non-linear inter-dependencies between microbes and diseases, directly alleviating the limitations of traditional linear matrix factorization methods.

The remainder of this paper is structured as follows: Section 2 summarizes the related work on microbe-disease association prediction. Section 3 introduces the datasets, similarity construction, and the proposed DDHMDA framework in detail. Section 4 presents the experimental settings, performance comparisons, ablation studies, and case study results. Finally, Section 5 discusses the main findings, limitations, and future research directions.

2. Related Work

MDA prediction models can be mainly divided into three categories: network propagation-based methods, matrix factorization, and completion-based methods, and machine learning-based methods.

2.1. Network Propagation-Based Methods

Network propagation-based methods infer potential MDA by modeling the topological relationships between nodes in a heterogeneous microbe-disease network and utilizing network structures or propagation mechanisms of path information. Chen et al. [17] proposed a KATZ centrality metric-based method called KATZ measure for Human Microbe-Disease Association prediction (KATZHMDA) to obtain the potential MDA scores by calculating the weighted sum of all possible paths between microbe nodes and disease nodes. Subsequently, the Path-Based Human Microbe-Disease Association (PBHMDA) prediction method proposed by Huang et al. [18] further adopts a path enumeration strategy, systematically traversing various types of path structures in the heterogeneous network using depth-first search, and scoring unknown microbe-disease pairs based on path lengths and node weights. Guan et al. [19] proposed the Bi-Directional Heterogeneous Network Selection (BDHNS) model by introducing a bi-directional heterogeneous network selection mechanism, which performs random walk processes on the microbe network and the disease network, respectively, to achieve bi-directional propagation and information fusion, thereby improving the stability and robustness of potential association prediction.

2.2. Matrix Factorization and Completion-Based Methods

Matrix factorization and completion-based methods infer unknown associations by modeling the MDA prediction as a low-rank matrix completion problem from an overall structural perspective. For instance, He et al. [20] introduced the Graph Regularized Non-negative Matrix Factorization for Human Microbe-Disease Association (GRNMFHMDA) prediction model by adding a graph regularization constraint term into the non-negative matrix factorization framework, enabling the representations of microbes and diseases in the latent space to preserve both the original association structure and the geometric properties of similarity networks. The Matrix Completion for Human Microbe-Disease Associations (MCHMDA) method proposed by Yan et al. [21] is based on low-rank matrix completion theory and incorporates a fast singular value thresholding algorithm to reconstruct the association matrix, thereby effectively alleviating the modeling difficulties caused by the highly sparse nature of microbe-disease data. Liu et al. [22] proposed Minimizing Matrix Nuclear Norm for Microbe-Disease Association (MNNMDA) prediction by formalizing the MDA prediction problem as a matrix completion model with nuclear norm minimization and introducing a sparse regularization term to enhance robustness against noisy data.

2.3. Machine Learning-Based Methods

Machine learning-based methods typically model MDA prediction as a supervised learning problem, utilizing known MDAs as labeled samples to learn the mapping relationships between microbes and diseases in the feature space, and subsequently predicting potential MDA. For example, Wang et al. [23] proposed Laplacian Regularized Least Squares for Human Microbe-Disease Association (LRLSHMDA) via building semi-supervised classifiers in the microbe and disease space based on the Laplacian regularized least squares framework, and inferred potential MDAs by fusing the prediction results of the two spaces. By introducing graph Laplacian regularization terms, this method enables the model to fully utilize the structural information of microbe and disease similarity networks during the prediction process, thereby explicitly depicting the topological structure of the similarity network while leveraging the known association information. Long et al. [24] put forward the GATMDA model by introducing graph attention networks to perform end-to-end representation learning on the heterogeneous network composed of microbes and diseases, and combining an inductive matrix completion to model the interactive relationships between nodes. GATMDA utilizes a multi-head attention mechanism to perform weighted aggregation of neighbor nodes, effectively enhancing node representations, and realizes the prediction of unknown associations through an inductive matrix completion method. Furthermore, Peng et al. [25] presented the graph attention autoencoder, positive-unlabeled learning, and deep neural network for Microbe-Disease Association (GPUDDMDA) prediction method by integrating a graph attention autoencoder, positive-unlabeled (PU) learning strategies, and deep neural network structures to enhance the feature representation capabilities of microbe-disease pairs from multiple perspectives. By jointly modeling graph structure information and sample annotation information, this method achieves a more comprehensive characterization of potential association relationships and introduces deep non-linear mapping in the prediction stage to improve the model's expressive capabilities. The GCDSAEMDA model introduced by Wang et al. [26] performs weighted aggregation on heterogeneous neighbor nodes through the graph attention mechanism, and utilizes the sparse autoencoder to extract high-order latent features. Then, the heterogeneous graph attention network is combined with the deep sparse autoencoder for joint representation learning of multi-source features. Finally, an ensemble learning algorithm is employed to classify and predict the potential MDAs. Recent MDA models further combine heteroge-

neous graph learning, hypergraph convolution, and contrastive representation learning. DCGHCN jointly exploits graph and hypergraph channels to capture complementary structural information [27]; DMCFMDA performs dual-channel multi-source cross-modal fusion with contrastive learning [28]; and GRNCFMDA combines heterogeneous graph random attention with neural collaborative filtering [29]. Together with GCDSAEMDA and the PU-learning-based GPUDMDA, these methods provide a broader set of contemporary baselines for evaluating DDHMDA.

3. Materials and Methods

3.1. Human Microbe-Disease Associations

In this study, two public databases, HMDAD [7] and Disbiome [30], were utilized for model evaluation. HMDAD is a relatively small-scale dataset and its original data contains 39 diseases, 292 microbes, and 483 microbe-disease associations. After removing duplicates, 450 association pairs are retained. Disbiome is a larger-scale database, encompassing 375 diseases, 1615 microbes, and 10,866 association pairs. After duplicate removal, 8697 association pairs consisting of 1584 microbes and 355 diseases were finally obtained (see Table 1). There are obvious differences in scale and complexity between the two databases. The former is suitable for verifying the stability of the model in relatively sparse scenarios with fewer samples, while the latter can test the generalization ability and robustness of the model in large-scale and highly heterogeneous scenarios. Therefore, combining a large and a small dataset allows for a more comprehensive evaluation of the predictive performance of the proposed model from the perspectives of different data distributions and network structure characteristics.

Table 1. Statistics of the microbe-disease association databases.

Datasets	Microbes	Diseases	Associations
HMDAD	292	39	450
Disbiome	1584	355	8697

Based on the above association pair data, we represent the MDAs as an association matrix $Y \in \mathbb{R}^{N_d \times N_m}$, where the numbers of diseases and microbes are N_d and N_m , respectively. Elements with a value of 1 in the association matrix correspond to known verified associations, while elements with a value of 0 represent potential unconfirmed associations, which can be specifically expressed as:

$$Y_{ij} = \begin{cases} 1, & \text{if there is an association between disease } i \text{ and microbe } j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

3.2. Semantic Similarity of Diseases

The semantic relationships between diseases are modeled based on a Directed Acyclic Graph (DAG) constructed by the Disease Ontology (DO). Referring to the method proposed by Wang et al. [31], the hierarchical structure of diseases in the DAG is utilized to calculate the semantic contribution values of diseases, thereby characterizing the semantic similarity between diseases. Specifically, for any disease d , the semantic contribution of node x to disease d is defined as follows:

$$C_d(x) = \begin{cases} 1, & x = d \\ \max\{\lambda \cdot C_d(y) | y \in \text{child}(x)\}, & x \neq d \end{cases} \quad (2)$$

where λ is the semantic decay coefficient (it is set at 0.5), used to control the degree of decay when semantic information propagates upwards along the hierarchical structure.

The semantic value of disease d is defined as the cumulative sum of the semantic contribution values of its node and all its ancestor nodes, namely:

$$S(d) = \sum_{x \in \mathcal{P}(d)} C_d(x) \quad (3)$$

where $\mathcal{P}(d)$ represents the node set of disease d in the DAG, including d itself and all its ancestor nodes.

Based on the above definition, the semantic similarity between any two diseases d_p and d_q is calculated as follows:

$$\text{Sim}_{\text{sem}}(d_p, d_q) = \frac{\sum_{x \in \mathcal{P}(d_p) \cap \mathcal{P}(d_q)} (C_{d_p}(x) + C_{d_q}(x))}{S(d_p) + S(d_q)} \quad (4)$$

3.3. Functional Similarity of Microbes

The functional similarity of microbes is based on the assumption that microbes with similar association patterns at the disease level tend to share higher functional similarities [32]. Accordingly, the functional similarity of microbes is quantified through the semantic relationships between their associated disease sets. Let the sets of diseases associated with microbes m_u and m_v be respectively:

$$\mathcal{D}_u = \{d_{u1}, d_{u2}, \dots, d_{u\alpha}\} \quad (5)$$

$$\mathcal{D}_v = \{d_{v1}, d_{v2}, \dots, d_{v\beta}\} \quad (6)$$

For any disease $d_{ui} \in \mathcal{D}_u$, its similarity with the disease set $\mathcal{D}_v$ is defined as:

$$H(d_{ui}, \mathcal{D}_v) = \max_{d \in \mathcal{D}_v} \text{Sim}_{\text{sem}}(d_{ui}, d) \quad (7)$$

where $\text{Sim}_{\text{sem}}(\cdot, \cdot)$ denotes the semantic similarity between diseases.

On this basis, the functional similarity between microbes m_u and m_v is defined as the average of the similarities between the two disease sets:

$$\text{FS}(m_u, m_v) = \frac{\sum_{i=1}^{\alpha} H(d_{ui}, \mathcal{D}_v) + \sum_{j=1}^{\beta} H(d_{vj}, \mathcal{D}_u)}{\alpha + \beta} \quad (8)$$

3.4. GIP Similarity of Diseases and Microbes

In this study, the Gaussian Interaction Profile (GIP) kernel function is adopted to measure the similarity between microbes and diseases. This method transforms the interactive differences between entities into a quantifiable similarity measure by mapping association patterns into a Gaussian kernel space. Let the association matrix be Y , where the row vectors and column vectors represent the interaction features of microbes and diseases, respectively. For any two diseases d_p and d_q , their GIP similarity is defined as:

$$K_d(d_p, d_q) = \exp(-\eta_d \|\mathbf{y}_{\cdot p} - \mathbf{y}_{\cdot q}\|^2) \quad (9)$$

where $\mathbf{y}_{\cdot p}$ represents the column vector of the disease d_p in the association matrix.

The bandwidth parameter in the kernel function is adaptively determined by the overall scale of the interaction features, specifically defined as:

$$\eta_d = \frac{\kappa_d}{\frac{1}{N_d} \sum_{p=1}^{N_d} \|\mathbf{y}_{\cdot p}\|^2} \quad (10)$$

where N_d represents the number of diseases. κ_d is a normalization coefficient, which is set to 1 in this study.

Similarly, for any two microbes m_u and m_v , their GIP similarity is calculated as:

$$K_m(m_u, m_v) = \exp(-\eta_m \|\mathbf{y}_{u\cdot} - \mathbf{y}_{v\cdot}\|^2) \quad (11)$$

where $\mathbf{y}_{u\cdot}$ represents the row vector of the microbe m_u in the association matrix. The kernel bandwidth parameter is defined as:

$$\eta_m = \frac{\kappa_m}{\frac{1}{N_m} \sum_{u=1}^{N_m} \|\mathbf{y}_{u\cdot}\|^2} \quad (12)$$

where N_m represents the number of microbes and κ_m is a normalization coefficient, which is set to 1 in this paper.

3.5. Similarity Integration

Given that disease semantic similarity and microbe functional similarity generally suffer from sparsity issues in real-world data, this study uniformly integrates multi-source similarity information to obtain more robust feature representations of disease and microbe similarity. Specifically, we respectively integrate the semantic/functional similarity with the GIP similarity to construct a comprehensive similarity matrix.

For a disease pair (d_p, d_q) , the comprehensive similarity is defined as:

$$S_d = \tilde{S}_d(d_p, d_q) = \begin{cases} \frac{1}{2}(\text{Sim}_{\text{sem}}(d_p, d_q) + K_d(d_p, d_q)), & \text{Sim}_{\text{sem}}(d_p, d_q) \neq 0 \\ K_d(d_p, d_q), & \text{otherwise} \end{cases} \quad (13)$$

Similarly, for a microbe pair (m_u, m_v) , the integrated similarity is defined as:

$$S_m = \tilde{S}_m(m_u, m_v) = \begin{cases} \frac{1}{2}(\text{FS}(m_u, m_v) + K_m(m_u, m_v)), & \text{FS}(m_u, m_v) \neq 0 \\ K_m(m_u, m_v), & \text{otherwise} \end{cases} \quad (14)$$

3.6. DDHMDA

In this paper, we propose DDHMDA for predicting the potential human microbe-disease associations. DDHMDA consists of four parts: (1) local feature encoding; (2) dynamic hypergraph construction; (3) hypergraph convolution module; (4) prediction module. The overall framework of DDHMDA is illustrated in Figure 1, and the detailed implementation of each module is presented as follows.

(1) Local feature encoding

Initially, to obtain the local representations of microbes and diseases, we input the integrated similarity of microbes and diseases, namely $S_m \in \mathbb{R}^{N_m \times N_m}$ and $S_d \in \mathbb{R}^{N_d \times N_d}$, into a linear transformation and GCN module. Specifically, we first construct a K -Nearest Neighbor (KNN)-based topological graph for both microbes and diseases based on their respective integrated similarity matrices. To ensure topological symmetry and numerical stability, the adjacency matrices are symmetrized, augmented with self-loops, and subsequently row-normalized to obtain the normalized adjacency matrix A_{norm}^d and A_{norm}^m . For the initialization of node attributes, we adopt a row-based linear projection strategy. For

either the microbe or disease modality, its initial similarity matrix (generically denoted as S , which corresponds to S_m or S_d) is directly utilized as the initial node attributes. It is then projected into a low-dimensional embedding space via a linear transformation:

$$X^{(0)} = \text{LayerNorm}(\text{Dropout}(SW_{\text{proj}} + \mathbf{b})) \quad (15)$$

where $S \in \mathbb{R}^{N \times N}$ is the input similarity matrix, $W_{\text{proj}} \in \mathbb{R}^{N \times d}$ is a learnable weight matrix, $\mathbf{b}$ is the bias term, and N and d represent the dimension of the similarity matrix and the embedding.

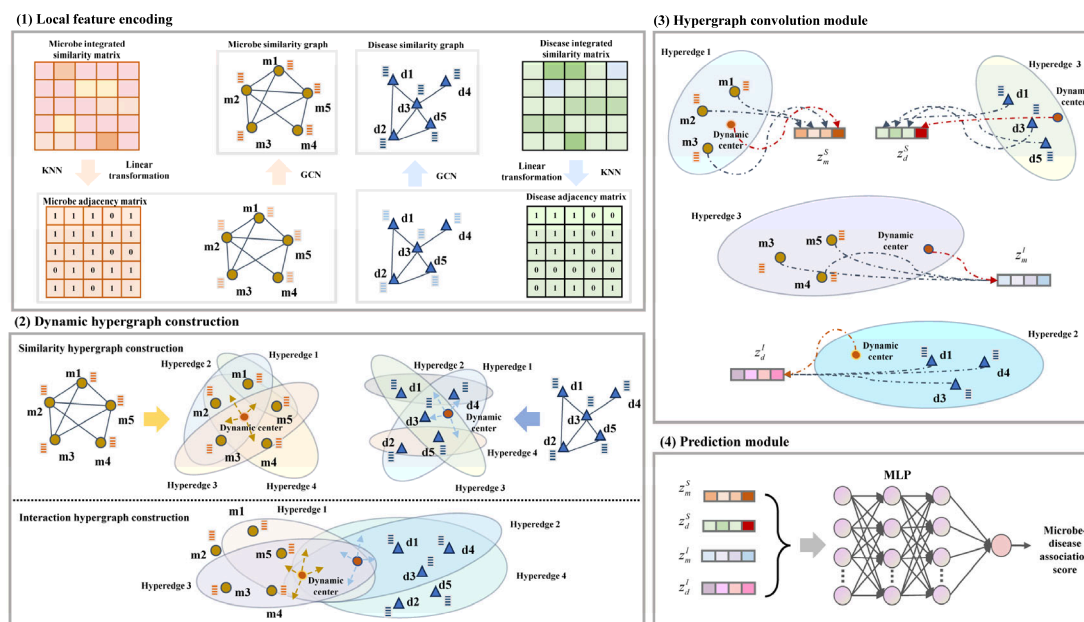


Figure 1. Overall architecture of DDHMDA. Fold-specific integrated microbe and disease similarities are first encoded by local KNN-GCN branches. A differentiable K-means similarity hypergraph and an attention-based interaction hypergraph then capture intra-modal and cross-modal high-order structure, respectively. Their embeddings are fused for microbe-disease association prediction.

Subsequently, a multi-layer GCN is employed to capture local topological information [29]. To mitigate the over-smoothing problem and stabilize the training process, we introduce L_2 norm normalization and residual connections during message passing. The message passing at the l -th layer is defined as:

$$M^{(l)} = \frac{A_{\text{norm}} X^{(l-1)}}{\|A_{\text{norm}} X^{(l-1)}\|_2} \quad (16)$$

$$X^{(l)} = \text{Dropout}(X^{(l-1)} + M^{(l)}) \quad (17)$$

where the normalized adjacency matrix A_{norm}^d and A_{norm}^m are generically denoted as A_{norm} .

After L layers of convolution, the final local node representation H is aggregated by averaging the initial features and the messages from all layers:

$$H = \frac{1}{L+1} \left(X^{(0)} + \sum_{l=1}^L M^{(l)} \right) \quad (18)$$

Through this encoding module, we obtain the local node embeddings $H_m \in \mathbb{R}^{n_m \times d}$ and $H_d \in \mathbb{R}^{n_d \times d}$ for microbes and diseases, respectively.

(2) Dynamic hypergraph construction

Inspired by the paper [33,34], we construct a dual dynamic hypergraph module to capture intra-modal global similarities and inter-modal interactions. For the similarity hypergraph, differentiable K-means clustering is used to learn high-order structure within each modality [35]. Let K denote the number of clusters, which also determines the number of dynamic hyperedges. Additionally, the model retains a set of learnable cluster centers $C \in \mathbb{R}^{K \times d}$. The cosine distance between node h_i and hyperedge center c_j is formulated as:

$$d(h_i, c_j) = 1 - \frac{h_i \cdot c_j}{\|h_i\|_2 \|c_j\|_2} \quad (19)$$

By introducing a temperature parameter β , we calculate the soft attention weights, namely assignment probabilities $R \in \mathbb{R}^{N \times K}$, for allocating nodes to various hyperedges:

$$r_{ij} = \frac{\exp(-\beta \cdot d(h_i, c_j))}{\sum_{k=1}^K \exp(-\beta \cdot d(h_i, c_k))} \quad (20)$$

During the forward process, the cluster centers C are dynamically differentiable updated using an Expectation-Maximization (EM)-like step:

$$c_k = \frac{\sum_{i=1}^N r_{ik} h_i}{\sum_{i=1}^N r_{ik}} \quad (21)$$

To enforce hypergraph sparsity and filter out noise, we apply a Top- t sparsification operation, preserving only the t largest assignment probabilities for each node while setting the rest to zero. This yields the final similarity hypergraph incidence matrices Y_m^S and Y_d^S , along with the updated centers C_m and C_d for microbes and diseases, respectively.

For interaction hypergraph construction, beyond intra-modal similarities, the synergistic cross-modal interactions between microbes and diseases are critical. We construct an interaction hypergraph where nodes from one modality are forced to attend to the hyperedge cluster centers of the other modality. Taking microbe nodes as an example, the interaction score matrix between the microbe node embeddings H_m and the disease cluster centers C_d is calculated as:

$$S_{int} = \sigma(H_m \mathbf{W}_{int} C_d^T) \quad (22)$$

where $\mathbf{W}_{int} \in \mathbb{R}^{d \times d}$ is a learnable bilinear interaction weight matrix and $\sigma(\cdot)$ denotes the Sigmoid activation function.

This is followed by a SoftMax normalization along the hyperedge dimension to derive the cross-modal association weights α :

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^K \exp(s_{ik})} \quad (23)$$

Similarly, the Top- t sparsification is applied to yield the interaction hypergraph incidence matrix for microbes, denoted as Y_m^I . Symmetrically, utilizing disease nodes H_d and microbe cluster centers C_m , we construct the interaction hypergraph incidence matrix for diseases Y_d^I .

(3) Hypergraph convolution module

After establishing both the similarity hypergraph Y^S and the interaction hypergraph Y^I , we employ a hypergraph convolutional network module designed for hyperedge feature aggregation and node feature updating.

First, node features are projected via a learnable weight matrix W_{hyper} and aggregated onto the hyperedges based on the incidence matrix Y , resulting in the aggregated feature $E_{agg} \in \mathbb{R}^{K \times d}$:

$$E_{agg} = (Y^T (H W_{hyper})) \quad (24)$$

Next, the cluster center representations C are concatenated with the aggregated features E_{agg} , and fed into a Multi-Layer Perceptron, namely, MLP_{edge} to generate the refined hyperedge embeddings E :

$$E = \text{Dropout}(MLP_{edge}([C \parallel E_{agg}])) \quad (25)$$

Finally, the hyperedge information is propagated back to the nodes using the incidence matrix Y . Combined with residual connections and Layer Normalization, the final node representations are updated as follows:

$$Z = \text{LayerNorm}(H + \text{ReLU}(YE)) \quad (26)$$

Through this convolutional block, the representations of microbes and diseases on both the similarity and interaction hypergraphs are updated to $Z_m^S, Z_m^I, Z_d^S, Z_d^I$, respectively.

(4) Prediction module

To comprehensively leverage the multi-scale and multi-modal features extracted above, the model concatenates the four sets of embeddings for a given microbe-disease pair (i, j) :

$$x_{ij} = [z_{m,i}^S \parallel z_{m,i}^I \parallel z_{d,j}^S \parallel z_{d,j}^I] \quad (27)$$

where $x_{ij} \in \mathbb{R}^{4d}$. Ultimately, the fused feature vector is fed into a prediction module, which consists of fully connected layers, ReLU activations, and dropout layers, to output the predicted association score for the microbe-disease pair:

$$\hat{y}_{ij} = MLP_{pred}(x_{ij}) \quad (28)$$

4. Results

4.1. Experimental Settings and Evaluation Metrics

All experiments were implemented in Python 3.9.25 (Python Software Foundation, Beaverton, OR, USA) using the PyTorch framework version 2.2.0 with CUDA 11.8 support (PyTorch Foundation, San Francisco, CA, USA; NVIDIA Corporation, Santa Clara, CA, USA). The final DDHMDA configuration retained five nearest neighbors for local KNN graph construction and used an embedding dimension of 256, five GCN layers, dropout of 0, eight clusters (and thus eight dynamic similarity hyperedges), temperature parameter $\beta = 2$, Top- $t = 3$, three differentiable K-means update iterations, a hypergraph hidden dimension of 256, a predictor hidden dimension of 128, a learning rate of 1×10^{-4} , weight decay of 5×10^{-4} , and a maximum of 200 epochs. Training used full-batch Adam optimization and a validation ratio of 0.1. In the method formulation, K denotes the number of differentiable K-means clusters and corresponding dynamic hyperedges. By contrast, the numerical suffixes in Static_Sim_HG_K5, Static_Sim_HG_K10, and Static_Sim_HG_K15 identify the fixed neighborhood sizes used only in the static-hypergraph ablation.

To prevent information leakage, every fold is processed independently. Known positive associations are first divided into training, validation, and test subsets. A fold-specific training association matrix is then constructed exclusively from the training positives.

Disease GIP similarity, microbe GIP similarity, microbe functional similarity, and the integrated similarities are recomputed from this training association matrix within the fold. DO semantic similarity is an external prior independent of the association split, but its integration with the fold-specific disease GIP kernel is performed after splitting. The local KNN graphs and both dynamic hypergraphs are then built from these fold-specific training features. Validation and test associations are never used to construct similarities, graphs, hypergraphs, or training inputs.

All unobserved microbe-disease pairs, after excluding every known positive in the full database, form the candidate-negative pool. To reduce the distributional bias of purely random sampling, we employed a cluster-based representative negative-sampling strategy combining K-means clustering and cosine-distance ranking (denoted as K-means/cosine). In each fold, candidate unobserved pairs were represented by concatenating the fold-specific feature vectors of the corresponding microbe and disease. The candidate pairs were partitioned into eight clusters using K-means and ranked within each cluster according to their cosine distances from the corresponding cluster center. Pairs closer to the cluster centers were preferentially selected as representative negative samples. This procedure aims to improve the representativeness of the sampled negative set rather than to assume that the selected pairs are experimentally confirmed negatives. Furthermore, training, validation, and test negatives are mutually disjointed and sampled in a 1:1 ratio to their corresponding positives. For CV1 and CV2, negative candidates are restricted to the appropriate training, validation, or held-out disease/microbe entities; for CV3, one disjoint candidate set is selected and partitioned according to the positive split sizes. Negative sets are fixed within each fold and sampling repetition and are not resampled between epochs. We repeat evaluation with ten negative-sampling seeds (1–10) and report the mean $\pm$ standard deviation across 50 fold-seed runs. Because unobserved pairs may contain undiscovered positives, this remaining label uncertainty is explicitly acknowledged as a limitation. Furthermore, hyperparameters were selected using validation data only. The search ranges were dropout in {0, 0.1, 0.2, 0.3, 0.5}, local KNN neighborhood size in {1, 3, 5, 10, 15}, GCN layers in {1, 2, 3, 4, 5}, number of differentiable K-means clusters and corresponding dynamic hyperedges in {8, 16, 32, 64}, Top- t in {1, 3, 5, 10, 15}, β in {0.5, 1, 2, 3, 5, 10}. The selected setting was subsequently examined by sensitivity analysis on HMDAD-CV3.

DDHMDA is optimized with binary cross-entropy with logits loss. The best checkpoint in each fold is selected by validation AUC. Training stops when validation AUC fails to improve by at least 1×10^{-6} for 20 consecutive epochs, and the best checkpoint is restored for testing. The decision threshold used to calculate ACC and F1 is not tuned on the test set; it is selected separately in each fold by maximizing F1 on the validation precision-recall curve and is then applied unchanged to the corresponding test predictions.

Moreover, this paper employs four classification evaluation metrics containing AUPR, AUC, ACC, and F1-score, to comprehensively evaluate the performance of the model. Among them, AUC is used to measure the model's overall ability to distinguish between positive and negative samples under different discrimination thresholds. It is defined as the area under the ROC curve (with False Positive Rate, FPR, as the horizontal axis and True Positive Rate, TPR, as the vertical axis), where:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \quad (29)$$

where TP , TN , FP , and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

AUPR evaluates the model's ability to identify positive samples by calculating the area under the Precision-Recall curve, where:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN} \quad (30)$$

Accuracy (ACC) represents the proportion of samples correctly predicted by the model out of the total samples, defined as:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (31)$$

The F1-score is the harmonic mean of precision and recall, used to comprehensively measure the accuracy and completeness of the model in predicting positive samples:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (32)$$

4.2. Performance Evaluation

We use five-fold cross-validation with ten negative-sampling seeds to assess three distinct generalization scenarios. CV3 is reported as the main experiment because it corresponds to the standard pair-level MDA prioritization task and permits direct comparison across methods. CV1 and CV2 impose stricter cold-start conditions for unseen diseases and microbes, respectively, and are therefore reported as supplementary analyses in Appendix A. Figure 2 summarizes the three split strategies.

- CV1 (disease-level cold start, Appendix A): diseases are divided into five folds. Associations involving the held-out diseases are excluded from the training association matrix and used only for testing, while a validation subset is selected from the training-side diseases. This setting evaluates prediction for diseases with no training associations.
- CV2 (microbe-level cold start, Appendix A): microbes are divided into five folds. Associations involving the held-out microbes are excluded from the training association matrix and used only for testing, while a validation subset is selected from the training-side microbes. This setting evaluates prediction for previously unseen microbes.
- CV3 (pair-level CV, main setting): known positive microbe-disease pairs are divided into five folds. In each fold, 20% of the positive pairs are held out for testing, and 10% of the remaining training-side positives are used for validation. All association-derived similarities, graphs, and hypergraphs are reconstructed from the remaining training positives only.

To determine whether the proposed high-order modeling strategy provides an advantage over representative existing approaches, we compared DDHMDA with seven methods on HMDAD and Disbiome under the main CV3 setting: DCGHCN [27], DM-CFMDA [28], GCDSAEMDA [26], GPUDMDA [25], GRNCFMDA [29], KATZHMDA [17], and MNNMDA [22]. This set covers path-based prediction, matrix completion, PU learning, heterogeneous graph neural networks, graph-hypergraph learning, and contrastive cross-modal fusion. For a fair comparison, all methods use the same positive folds and sampled training, validation, and test pairs; association-derived inputs are reconstructed from the corresponding training association matrix where required. The comparative results are summarized in Table 2.

As shown in Table 2, DDHMDA achieves the best overall CV3 results on both datasets. On HMDAD, it obtains $\text{AUC} = 0.9789 \pm 0.0177$, $\text{AUPR} = 0.9843 \pm 0.0129$, $\text{ACC} = 0.9273 \pm 0.0423$, and $\text{F1} = 0.9217 \pm 0.0511$. Relative to KATZHMDA, the strongest baseline, the improvements are 0.0164 in AUC, 0.0098 in AUPR, 0.0235 in ACC, and 0.0224

in F1. KATZHMDA also performs strongly on this pair-level task, indicating that short-path information remains informative; nevertheless, the consistent gains of DDHMDA across both ranking-based and threshold-dependent metrics suggest that combining local topology with adaptive high-order relationships provides additional discriminative information. Several neural baselines show substantially larger standard deviations on the small and sparse HMDAD dataset, which suggests greater sensitivity to limited training associations and fold composition.

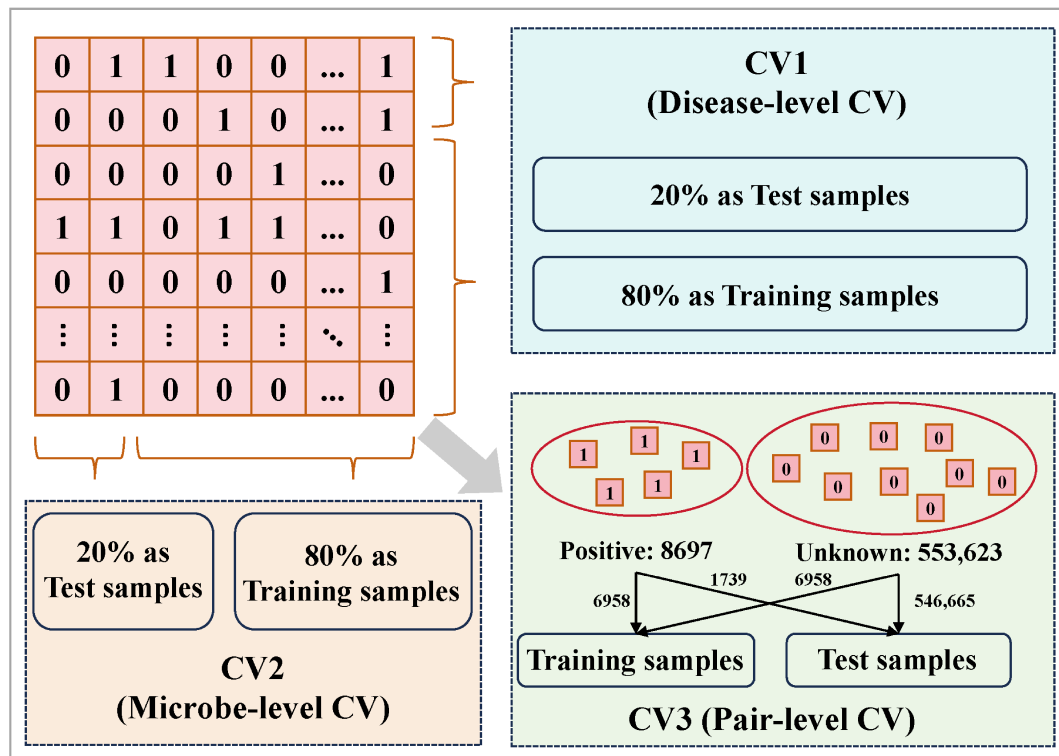


Figure 2. Data partitioning under the three evaluation settings. CV1 holds out complete disease nodes, CV2 holds out complete microbe nodes, and CV3 holds out association pairs. In every setting, validation and test positives, together with their sampled negatives, are excluded from the training association matrix before fold-specific similarities and graph and hypergraph structures are constructed.

On the larger Disbiome dataset, DDHMDA reaches $AUC = 0.9651 \pm 0.0042$, $AUPR = 0.9740 \pm 0.0031$, $ACC = 0.9226 \pm 0.0077$, and $F1 = 0.9194 \pm 0.0080$. Compared with KATZHMDA, DDHMDA improves AUC and AUPR by 0.0206 and 0.0169, respectively, while the corresponding ACC and F1 gains are 0.0392 and 0.0419. The recent neural baselines generally perform better on Disbiome than on HMDAD, which is consistent with their ability to benefit from a larger association network. DDHMDA nevertheless retains the highest values for all four metrics and exhibits small standard deviations, indicating that its advantage is maintained across dataset scales and is not restricted to a particular decision threshold.

Figure 3 provides a threshold-independent view of the ranking behavior summarized in Table 2. The ROC and precision-recall curves show that DDHMDA maintains the strongest overall discrimination on both the small, sparse HMDAD network and the larger Disbiome network. The advantage is particularly evident in the HMDAD precision-recall curve, indicating improved prioritization of positive associations when known links are limited. On Disbiome, the curves of several recent baselines become more competitive, but DDHMDA retains the leading AUC and AUPR. Together with the ACC and F1 results,

these observations indicate that the improvement is consistent across ranking-based and threshold-dependent evaluation criteria.

Table 2. Comparative performance under the main CV3 setting. Values are reported as mean $\pm$ standard deviation across the five folds and ten negative-sampling seeds.

Datasets	Method	AUC	AUPR	ACC	F1
HMDAD	DCGHCN	0.6451 $\pm$ 0.1036	0.6778 $\pm$ 0.0680	0.6087 $\pm$ 0.1208	0.6867 $\pm$ 0.0677
	DMCFMDA	0.7793 $\pm$ 0.0850	0.7683 $\pm$ 0.0764	0.7610 $\pm$ 0.0807	0.7862 $\pm$ 0.0696
	GCDSAEMDA	0.6079 $\pm$ 0.1312	0.6660 $\pm$ 0.1075	0.6339 $\pm$ 0.1013	0.7183 $\pm$ 0.0506
	GPUDMDA	0.5198 $\pm$ 0.1568	0.5391 $\pm$ 0.1114	0.6006 $\pm$ 0.0900	0.6963 $\pm$ 0.0534
	GRNCFMDA	0.7292 $\pm$ 0.1477	0.7451 $\pm$ 0.1079	0.7189 $\pm$ 0.1212	0.7642 $\pm$ 0.0737
	KATZHMDA	0.9625 $\pm$ 0.0128	0.9745 $\pm$ 0.0079	0.9038 $\pm$ 0.0257	0.8993 $\pm$ 0.0314
	MNNMDA	0.8605 $\pm$ 0.0550	0.8502 $\pm$ 0.0634	0.8284 $\pm$ 0.0562	0.8411 $\pm$ 0.0467
	DDHMDA	0.9789 $\pm$ 0.0177	0.9843 $\pm$ 0.0129	0.9273 $\pm$ 0.0423	0.9217 $\pm$ 0.0511
Disbiome	DCGHCN	0.8776 $\pm$ 0.0140	0.8985 $\pm$ 0.0116	0.8066 $\pm$ 0.0183	0.8051 $\pm$ 0.0156
	DMCFMDA	0.9126 $\pm$ 0.0177	0.9295 $\pm$ 0.0150	0.8363 $\pm$ 0.0220	0.8393 $\pm$ 0.0203
	GCDSAEMDA	0.7919 $\pm$ 0.0473	0.7830 $\pm$ 0.0594	0.7070 $\pm$ 0.0521	0.7507 $\pm$ 0.0268
	GPUDMDA	0.7854 $\pm$ 0.0213	0.7405 $\pm$ 0.0273	0.7640 $\pm$ 0.0209	0.7802 $\pm$ 0.0166
	GRNCFMDA	0.8593 $\pm$ 0.0301	0.8804 $\pm$ 0.0263	0.7963 $\pm$ 0.0238	0.7998 $\pm$ 0.0237
	KATZHMDA	0.9445 $\pm$ 0.0077	0.9571 $\pm$ 0.0067	0.8834 $\pm$ 0.0124	0.8775 $\pm$ 0.0121
	MNNMDA	0.9034 $\pm$ 0.0115	0.9280 $\pm$ 0.0091	0.8432 $\pm$ 0.0145	0.8313 $\pm$ 0.0144
	DDHMDA	0.9651 $\pm$ 0.0042	0.9740 $\pm$ 0.0031	0.9226 $\pm$ 0.0077	0.9194 $\pm$ 0.0080

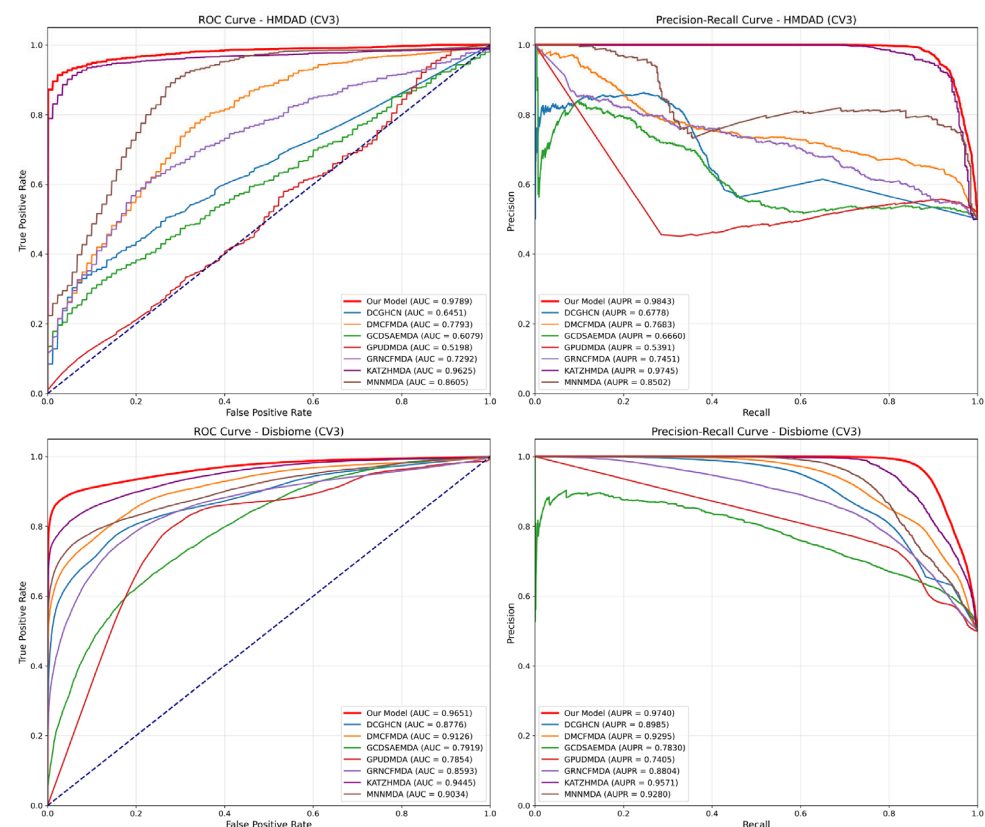


Figure 3. ROC and precision-recall curves for DDHMDA and seven baselines under CV3. The upper panels show HMDAD and the lower panels show Disbiome. Curves aggregate the held-out predictions over five folds and ten negative-sampling seeds; the label ‘Our Model’ in the exported plots denotes DDHMDA.

To determine whether the observed improvements could be attributed to variation across matched evaluations, two-sided paired *t*-tests were conducted for HMDAD-CV3 us-

ing the fold- and negative-seed-matched results of DDHMDA and each baseline. DDHMDA was significantly better than every baseline for both AUC and AUPR (all p -values < 0.001). Even against KATZHMDA, the closest competitor, the paired t -test yielded $p = 7.12 \times 10^{-10}$ for AUC and $p = 2.61 \times 10^{-8}$ for AUPR. These results indicate that the performance gains are consistent across the matched fold-seed evaluations and support the specific conclusion that DDHMDA achieved the best performance among the methods examined under the leakage-free CV3 protocol.

Moreover, the supplementary cold-start experiments further assess generalization to unseen entities. Under CV1, DDHMDA achieved the highest AUC and AUPR on HMDAD and the best overall performance on Disbiome. Under CV2, DDHMDA remained competitive with KATZHMDA on HMDAD and achieved the best performance across all four metrics on Disbiome. Overall, the lower performance under CV1 and CV2 than under CV3 indicates that predicting associations for completely unseen diseases or microbes remains challenging because their association-derived profiles contain limited information. Detailed results are provided in Figures A1 and A2 as well as Tables A1 and A2 in Appendix A.

To examine optimization convergence and assess whether prolonged training leads to overfitting, we recorded the training and validation binary cross-entropy losses across epochs on HMDAD and Disbiome under the CV3 setting and averaged them over the fold-seed evaluations. As shown in Figure 4, both training and validation losses decrease rapidly during the early epochs, demonstrating that the model learns the main association patterns efficiently. The validation loss subsequently reaches a plateau and fluctuates within a relatively narrow range, while the training loss continues to decrease. This divergence indicates that further optimization mainly improves the fit to the training pairs without a corresponding validation benefit. Selecting the checkpoint with the highest validation AUC and applying a patience of 20 therefore prevents later epochs from determining the test model and reduces overfitting.

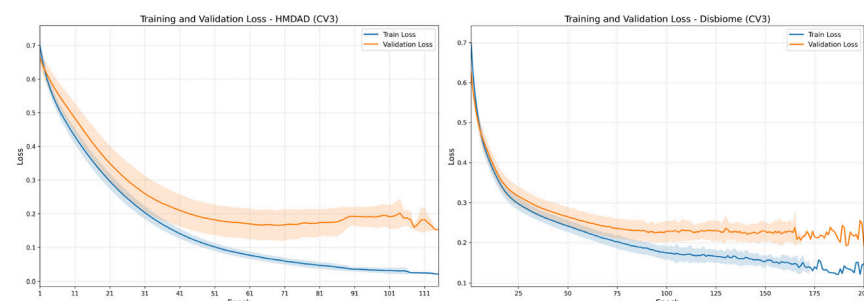


Figure 4. Mean training and validation binary cross-entropy loss on HMDAD and Disbiome under CV3. Solid lines show the fold-seed mean and the shaded regions represent the mean $\pm$ standard deviation. Validation loss decreases rapidly and then fluctuates within a narrow range, while early stopping restores the checkpoint with the highest validation AUC.

To evaluate whether the reported performance depends on narrowly tuned hyperparameter values, we varied the principal structural and optimization parameters individually around the selected configuration on HMDAD-CV3. As a result, Figure 5 shows that AUC and AUPR remain relatively stable around the selected configuration. Moderate changes in beta, dropout, GCN depth, local KNN neighborhood size, the number of dynamic hyperedges, and Top- t sparsification produce only limited variation in predictive performance. Performance deteriorates more noticeably only at some extreme settings, where excessive sparsification or model complexity can remove useful neighborhood information or increase optimization difficulty. These results indicate that the final performance is not produced by one isolated hyperparameter choice and support the robustness of the validation-selected configuration.

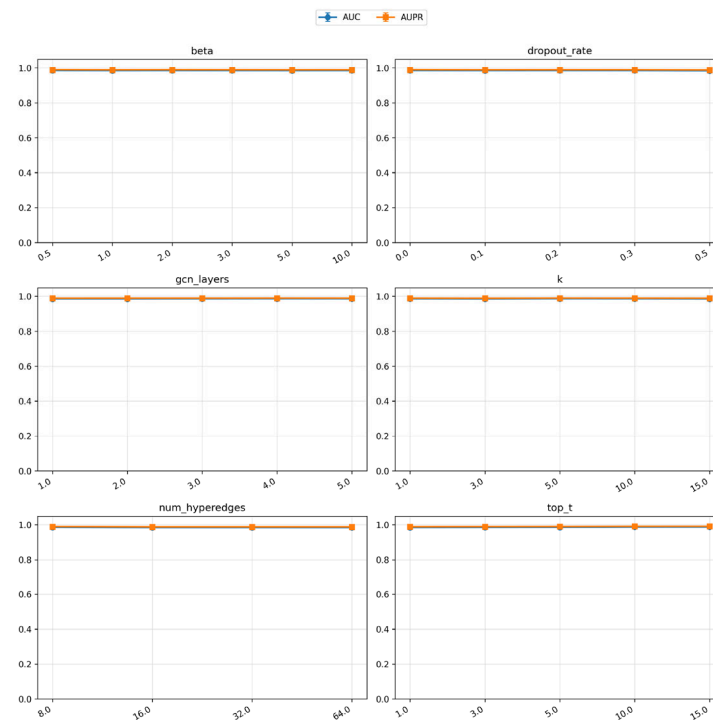


Figure 5. Hyperparameter sensitivity of DDHMDA on HMDAD-CV3. AUC and AUPR are shown while varying beta, dropout, GCN depth, local KNN neighborhood size, number of dynamic hyperedges, and Top- t sparsity; error bars summarize fold-seed variability.

4.3. Ablation Experiments

To thoroughly evaluate the contribution and necessity of each core component of DDHMDA, we conducted comprehensive ablation studies. Ablation experiments were performed on HMDAD under the main CV3 protocol using the same fold-wise feature reconstruction, ten negative-sampling seeds, validation threshold selection, and early stopping. In addition to removing architectural components, we compared random, plain K-means, and K-means/cosine negative-sampling strategies. Here, we designed the following degraded variants of our model by selectively removing or replacing specific modules:

- **No_Local_GCN:** Disables the GCN message-passing mechanism. The model bypasses local topological aggregation and relies solely on the linearly projected node attributes.
- **No_HG:** Completely removes the dual hypergraph learning branch. The prediction relies entirely on the local features extracted by the initial encoders.
- **No_Sim_HG:** Removes the differentiable K-means similarity hypergraph module, retaining only the cross-modal interaction hypergraph.
- **No_Int_HG:** Removes the attention-based cross-modal interaction hypergraph module, retaining only the similarity hypergraph.
- **Static_Sim_HG_K5/K10/K15:** Replaces the dynamic similarity hypergraph with fixed KNN-based hypergraphs whose neighborhood sizes are 5, 10, and 15, respectively. These suffix values refer only to the static-hypergraph ablation and are distinct from the five-neighbor setting of the local GCN graph.
- **Static_Int_HG:** Replaces the dynamic cross-modal attention mechanism with static associations. It directly utilizes the known training association matrix as the incidence matrix.

Furthermore, DDHMDA_Random and DDHMDA_Kmeans use random and plain K-means negative sampling, respectively; the complete DDHMDA uses K-means/cosine sampling. The performance of all variants is summarized in Table 3.

Table 3. Ablation study results on the HMDAD dataset under the CV3 setting.

Variants	AUC	AUPR	ACC	F1
DDHMDA_Random	0.8551 $\pm$ 0.0405	0.8566 $\pm$ 0.0403	0.7910 $\pm$ 0.0410	0.7863 $\pm$ 0.0462
DDHMDA_Kmeans	0.8520 $\pm$ 0.0426	0.8583 $\pm$ 0.0401	0.7786 $\pm$ 0.0539	0.7788 $\pm$ 0.0554
No_HG	0.8733 $\pm$ 0.0620	0.8896 $\pm$ 0.0652	0.8179 $\pm$ 0.0617	0.8243 $\pm$ 0.0534
No_Int_HG	0.9722 $\pm$ 0.0236	0.9789 $\pm$ 0.0182	0.9193 $\pm$ 0.0477	0.9147 $\pm$ 0.0538
No_Local_GCN	0.9785 $\pm$ 0.0194	0.9838 $\pm$ 0.0153	0.9261 $\pm$ 0.0479	0.9208 $\pm$ 0.0564
No_Sim_HG	0.9564 $\pm$ 0.0419	0.9673 $\pm$ 0.0316	0.9024 $\pm$ 0.0583	0.8954 $\pm$ 0.0646
Static_Int_HG	0.9759 $\pm$ 0.0100	0.9837 $\pm$ 0.0076	0.9306 $\pm$ 0.0433	0.9252 $\pm$ 0.0537
Static_Sim_HG_K5	0.9671 $\pm$ 0.0314	0.9755 $\pm$ 0.0227	0.9077 $\pm$ 0.0504	0.9037 $\pm$ 0.0526
Static_Sim_HG_K10	0.9691 $\pm$ 0.0294	0.9764 $\pm$ 0.0227	0.9134 $\pm$ 0.0435	0.9074 $\pm$ 0.0510
Static_Sim_HG_K15	0.9743 $\pm$ 0.0192	0.9805 $\pm$ 0.0145	0.9186 $\pm$ 0.0352	0.9135 $\pm$ 0.0424
DDHMDA	0.9789 $\pm$ 0.0177	0.9843 $\pm$ 0.0129	0.9273 $\pm$ 0.0423	0.9217 $\pm$ 0.0511

As illustrated in Table 3, the complete DDHMDA obtains the highest AUC (0.9789 $\pm$ 0.0177) and AUPR (0.9843 $\pm$ 0.0129). Removing the entire hypergraph branch produces the largest architectural decline, to AUC = 0.8733 and AUPR = 0.8896, confirming that high-order modeling provides information not recovered by the local encoder alone. Removing the similarity hypergraph causes a larger reduction than removing the interaction hypergraph, while both variants underperform the complete model, supporting their complementary roles. No_Local_GCN remains close to DDHMDA, indicating that the dynamic hypergraphs carry most of the ranking signal under CV3, although the full model still provides small improvements across all four metrics. The static similarity-hypergraph variants improve as their fixed neighborhood size increases from 5 to 15 but remain below DDHMDA in AUC and AUPR, supporting adaptive hyperedge learning. In the leakage-free rerun, Static_Sim_HG_K5 no longer exceeds the full model in ACC or F1. Static_Int_HG shows marginally higher ACC and F1, but these decision-threshold-dependent differences are small relative to their standard deviations and arise from fold-specific validation threshold selection; DDHMDA remains superior on the threshold-independent AUC and AUPR metrics. Finally, both DDHMDA_Random and DDHMDA_Kmeans are markedly weaker than the complete model, demonstrating that negative-sample selection materially affects performance estimates.

4.4. Case Studies

To further validate the practical capability of the DDHMDA framework in discovering unconfirmed MDAs in the HMDAD and Disbiome datasets, we conducted case studies on two complex diseases: asthma and IBD. We trained the model using all known associations in the two distinct datasets and ranked the unknown microbe-disease candidate pairs based on their predicted scores. The top 20 candidate microbes for each disease were subsequently verified through the existing biomedical literature.

4.4.1. Case Study on Asthma

Asthma is a prevalent chronic inflammatory airway disease, clinically characterized by recurrent episodes of coughing, chest tightness, wheezing, and dyspnea [36]. Recently, as microbiome research has advanced, accumulating evidence suggests that early colonization patterns of the respiratory and intestinal microecology are intricately linked to immune system development, potentially contributing to the pathogenesis of allergic diseases like asthma [4]. Particularly during the critical developmental window of infancy, structural alterations in the gut microbiota have been statistically associated with subsequent asthma risk, offering a novel theoretical framework for understanding asthma pathogenesis from a microecological perspective [37]. In this study, we present the top 20 asthma-associated candidate microbes predicted by the DDHMDA model based on the HMDAD and Disbiome

datasets, detailed in Tables 4 and 5, respectively. The tables outline the rank (Column 1), microbe name (Column 2), DDHMDA prediction score (Column 3), and validation evidence from relevant databases or literature (Column 4). If verified by a database, the database name is provided; if verified by literature, the corresponding PMID is listed.

Table 4. Top 20 candidate microbes associated with asthma predicted by DDHMDA on the HMDAD dataset.

Rank	Microbe Name	Prediction Score	Evidence
1	<i>Clostridium difficile</i>	0.9507	21872915
2	<i>Firmicutes</i>	0.9488	Unconfirmed
3	<i>Helicobacter pylori</i>	0.9451	HMDAD
4	<i>Clostridia</i>	0.9386	27634868
5	<i>Prevotella</i>	0.9358	HMDAD
6	<i>Bacteroidetes</i>	0.9316	HMDAD
7	<i>Proteobacteria</i>	0.8867	HMDAD
8	<i>Bifidobacterium</i>	0.8807	30290688
9	<i>Clostridium coccoides</i>	0.8666	21477358
10	<i>Prevotella copri</i>	0.8610	34310928
11	<i>Pseudomonas</i>	0.8564	25329665
12	<i>Lactobacillus</i>	0.8551	29479926
13	<i>Fusobacterium nucleatum</i>	0.8493	Unconfirmed
14	<i>Staphylococcus aureus</i>	0.8368	HMDAD
15	<i>Staphylococcus</i>	0.8193	HMDAD
16	<i>Actinobacteria</i>	0.7935	26220531
17	<i>Burkholderia</i>	0.7927	Unconfirmed
18	<i>Haemophilus</i>	0.7612	HMDAD
19	<i>Bacteroides</i>	0.7419	29241587
20	<i>Lachnospiraceae</i>	0.6901	39814146

Table 5. Top 20 candidate microbes associated with asthma predicted by DDHMDA on the Disbiome dataset.

Rank	Microbe Name	Prediction Score	Evidence
1	<i>Faecalibacterium</i>	0.9828	Disbiome
2	<i>Bacteroides</i>	0.9828	Disbiome
3	<i>Prevotella</i>	0.9815	Disbiome
4	<i>Lactobacillus</i>	0.9814	Disbiome
5	<i>Streptococcus</i>	0.9812	Disbiome
6	<i>Parabacteroides</i>	0.9805	Disbiome
7	<i>Blautia</i>	0.9803	Disbiome
8	<i>Roseburia</i>	0.9803	Disbiome
9	<i>Ruminococcus</i>	0.9800	Disbiome
10	<i>Bifidobacterium</i>	0.9799	Disbiome
11	<i>Veillonella</i>	0.9797	Disbiome
12	<i>Enterococcus</i>	0.9795	Disbiome
13	<i>Lachnospiraceae</i>	0.9791	39814146
14	<i>Clostridium</i>	0.9781	27634868
15	<i>Staphylococcus</i>	0.9775	Disbiome
16	<i>Dialister</i>	0.9770	Disbiome
17	<i>Acinetobacter</i>	0.9769	Disbiome
18	<i>Fusobacterium</i>	0.9769	Disbiome
19	<i>Dorea</i>	0.9768	39814146
20	<i>Enterobacteriaceae</i>	0.9765	Disbiome

Among the top 20 candidate microbes predicted using the HMDAD dataset (see Table 4), 17 have been successfully validated by existing databases and literature. Specif-

ically, 7 associations are directly corroborated by the HMDAD database, while the remaining 10 are supported by previous experimental studies. Among the top 20 candidate microbes predicted using the Disbiome dataset (see Table 5), all of them have been successfully verified through existing databases and literature. Specifically, 17 of the associations were directly confirmed by the Disbiome database, while the remaining 3 associations were supported by previous experimental studies. The specific evidence for the top 10 predicted microbes listed in Table 4 and those verified in the literature of Table 5 is as follows. Specifically, studies focusing on gut colonization in early infancy have demonstrated that the early colonization of *Clostridium difficile* is significantly associated with the subsequent development of recurrent wheeze, eczema, and doctor-diagnosed asthma [38]. A cohort study investigating the stool microbiome at 3 months of age and asthma in preschool children revealed that infants who later developed asthma exhibited a reduced relative abundance of *Lachnospira* and an increased relative abundance of *Clostridium neonatale*; notably, the *Lachnospira* or *Clostridium* ratio was negatively correlated with asthma risk [39]. In a study focusing on wheezing diseases in Chinese children, the fecal abundance of *Bifidobacterium* in pediatric patients was significantly lower than that in healthy controls, whereas no significant difference was observed in the abundance of *Lactobacillus* between the two groups [40]. Further analysis of denaturing gradient gel electrophoresis (DGGE) profiles of the neonatal intestinal microbiota showed a statistically significant association between microbial characteristics related to *Clostridium coccoides* and the subsequent development of asthma [40]. Furthermore, research on the maternal gut microbiome during pregnancy indicated that maternal carriage of *Prevotella copri* is associated with a decreased risk of immunoglobulin IgE-mediated food allergy, suggesting that the maternal microbiome may affect the occurrence of offspring allergic diseases containing asthma-related phenotypes through immunomodulatory pathways [41]. In infant cohort studies, the development of the gut microbiome was found to be associated with nocturnal cough, and this association was modified by asthma status, and the family *Lachnospiraceae* and its genera (such as *Dorea*) accounted for a considerable proportion of the significantly associated OTUs [39]. Consistently, reports on the relationship between *Clostridium* and preschool asthma further support its potential role in early gut microecology and asthma risk [39]. In Table 4, although the “unconfirmed” candidate microbes have not yet been directly validated by existing databases or literature, they may still have potential biological significance and provide useful clues for future experimental verification.

4.4.2. Case Study on IBD

IBD is a chronic, relapsing inflammatory disorder of the gastrointestinal tract, primarily encompassing Crohn’s disease and ulcerative colitis [42]. Accumulating evidence indicates that dysbiosis of the gut microbiota plays a crucial role in the pathogenesis, immune dysregulation, and progression of IBD [42,43]. To further demonstrate the predictive capability of our model, we applied DDHMDA to identify the top 20 candidate microbes associated with IBD using the HMDAD and Disbiome datasets, detailed in Tables 6 and 7, respectively. The tables outline the rank, microbe name, DDHMDA prediction score, and validation evidence.

Table 6. Top 20 candidate microbes associated with IBD predicted by DDHMDA on the HMDAD dataset.

Rank	Microbe Name	Prediction Score	Evidence
1	<i>Firmicutes</i>	0.8456	19235886
2	<i>Clostridium coccooides</i>	0.8311	27687331
3	<i>Bacteroidetes</i>	0.8223	27999802
4	<i>Prevotella</i>	0.8019	28542929
5	<i>Proteobacteria</i>	0.7752	HMDAD
6	<i>Clostridia</i>	0.7492	28506071
7	<i>Helicobacter pylori</i>	0.7489	24914359
8	<i>Clostridium difficile</i>	0.7329	22508484
9	<i>Bifidobacterium</i>	0.6491	34337079
10	<i>Bacteroides</i>	0.6413	12906096
11	<i>Lachnospiraceae</i>	0.6215	HMDAD
12	<i>Actinobacteria</i>	0.6127	HMDAD
13	<i>Haemophilus</i>	0.5564	33666710
14	<i>Lactobacillus</i>	0.5194	29854599
15	<i>Escherichia coli</i>	0.4920	23822920
16	<i>Veillonella</i>	0.4660	37836432
17	<i>Prevotella copri</i>	0.4489	28542929
18	<i>Bacteroides vulgatus</i>	0.4477	12906096
19	<i>Enterobacteriaceae</i>	0.4257	36268225
20	<i>Staphylococcus</i>	0.4153	31698044

Table 7. Top 20 candidate microbes associated with IBD predicted by DDHMDA on the Disbiome dataset.

Rank	Microbe Name	Prediction Score	Evidence
1	<i>Bacteroides</i>	0.9623	Disbiome
2	<i>Faecalibacterium</i>	0.9611	Disbiome
3	<i>Streptococcus</i>	0.9589	28618865
4	<i>Prevotella</i>	0.9579	Disbiome
5	<i>Parabacteroides</i>	0.9574	Disbiome
6	<i>Lactobacillus</i>	0.9571	29854599
7	<i>Veillonella</i>	0.9570	37836432
8	<i>Ruminococcus</i>	0.9568	Disbiome
9	<i>Roseburia</i>	0.9564	39166405
10	<i>Lachnospiraceae</i>	0.9563	17699621
11	<i>Blautia</i>	0.9562	Disbiome
12	<i>Bifidobacterium</i>	0.9559	Disbiome
13	<i>Enterococcus</i>	0.9556	Disbiome
14	<i>Enterobacteriaceae</i>	0.9540	Disbiome
15	<i>Clostridium</i>	0.9536	Unconfirmed
16	<i>Dialister</i>	0.9517	27390077
17	<i>Acinetobacter</i>	0.9496	Unconfirmed
18	<i>Dorea</i>	0.9494	36695274
19	<i>Faecalibacterium prausnitzii</i>	0.9492	Disbiome
20	<i>Haemophilus</i>	0.9491	33666710

Among the top 20 candidate microbes predicted on the HMDAD dataset (see Table 6), remarkably, all 20 have been successfully validated by existing databases or related literature. Specifically, 3 associations (*Proteobacteria*, *Lachnospiraceae*, and *Actinobacteria*) are directly confirmed by the HMDAD database, while the remaining 17 are robustly supported by independent experimental literature. Reviewing the literature-validated microbes within the top 5 yields compelling evidence. Based on the literature evidence

supporting the top 5 candidate microbes in Table 6, existing studies have substantiated the associations between several key microbial taxa and IBD activity or inflammatory phenotypes across multiple lines of evidence, including patient-based investigations, systematic reviews/meta-analyses, and mechanistic considerations. A qPCR-based study reported that, in patients with active IBD, the fecal abundance of the phylum *Firmicutes*, especially the *Clostridium leptum* and *Clostridium coccoides* groups, was reduced, along with a significantly decreased *Firmicutes* or *Bacteroidetes* ratio [44]. A systematic review and meta-analysis further demonstrated that *Clostridium coccoides* abundance was lower in active IBD than in remission [45]. Another meta-analysis integrating multiple studies indicated that the mean intestinal level of *Bacteroides* was significantly reduced in IBD patients compared with non-IBD controls, and was also lower during active disease than remission [46]. Moreover, a review emphasized that increased *Prevotella* is associated with enhanced Th17-related mucosal inflammation, and may promote inflammatory cell recruitment through induction of IL-8 and IL-6; animal studies likewise suggested that *Prevotella* colonization can induce or aggravate inflammatory phenotypes, supporting its potential link to intestinal mucosal inflammation in chronic disease contexts [47].

As detailed in Table 7, the predictive performance remains robust on the Disbiome dataset. Out of the top 20 predicted candidate microbes, 18 have been validated. Among these, 10 associations are explicitly confirmed by the Disbiome database records, and another 8 are substantiated by relevant experimental literature. Regarding the literature evidence for the top 5 candidate microbes in Table 7, a DSS-induced murine colitis study observed that intervention with *Streptococcus thermophilus* delayed colitis onset, mitigated clinical manifestations such as body-weight loss and gastrointestinal bleeding, and reduced bacterial translocation into colonic tissue, suggesting that *Streptococcus*-related strains may be involved in regulating IBD-associated inflammatory processes [48]. Furthermore, the “unconfirmed” candidate microbes may represent insufficiently explored microbe-disease associations and could serve as potential leads for subsequent validation of IBD-related microbes.

In summary, in the case studies of asthma and IBD, the vast majority of high-priority candidate microbes predicted by the DDHMDA model were strongly supported by database records or medical experimental literature. These reliable validation results demonstrate that the predicted associations generated by our model are based on a solid foundation, with high accuracy and reliability. This further indicates the performance of the DDHMDA framework in practical applications, suggesting that it can serve as a trustworthy computational auxiliary tool to accurately and reliably guide the clinical discovery and research of unknown MDAs.

5. Discussion

In this study, we developed a novel deep learning framework, DDHMDA, to effectively infer potential human microbe-disease associations. Its main contribution lies in jointly modeling local pairwise topology and high-order relationships. Specifically, local representations are first obtained from fold-specific similarity matrices using graph convolutional networks. A differentiable K-means similarity hypergraph is then employed to capture global intra-modal clustering patterns, while an attention-based interaction hypergraph models cross-modal relationships between microbes and diseases. This design enables the hypergraph structures to be optimized during training instead of relying entirely on predefined and fixed connections.

To avoid information leakage, all association-derived inputs were reconstructed independently within each cross-validation fold. Only the training positive associations were retained in the fold-specific association matrix. Disease GIP similarity, microbe GIP similar-

ity, microbe functional similarity, and the integrated similarity matrices were subsequently recomputed from this training matrix. The local KNN graphs and hypergraph inputs were then constructed from these fold-specific features. Validation and test associations were not used to construct similarities, graphs, hypergraphs, or training inputs. Under this leakage-free protocol, DDHMDA demonstrated consistent predictive performance across the CV1, CV2, and CV3 settings on both HMDAD and Disbiome. CV3 was used as the main evaluation setting because it represents the standard pair-level MDA prioritization task, whereas the more stringent disease-level and microbe-level cold-start results from CV1 and CV2 are reported in Appendix A.

As a result, under CV3, DDHMDA achieved AUC and AUPR values of 0.9789 ± 0.0177 and 0.9843 ± 0.0129 , respectively, on HMDAD. On Disbiome, it obtained an AUC of 0.9651 ± 0.0042 and an AUPR of 0.9740 ± 0.0031 . DDHMDA achieved the best overall CV3 performance among the eight evaluated methods, which included path-based prediction, matrix completion, positive-unlabeled learning, heterogeneous graph learning, graph-hypergraph convolution, contrastive cross-modal fusion, and neural collaborative filtering approaches. Compared with KATZHMDA, the strongest baseline under CV3, DDHMDA improved AUC and AUPR by 0.0164 and 0.0098 on HMDAD and by 0.0206 and 0.0169 on Disbiome, respectively. The paired analysis on HMDAD further indicated that DDHMDA significantly outperformed the evaluated baselines in terms of AUC and AUPR. These findings support the specific conclusion that DDHMDA performed best among the methods examined under the present leakage-free evaluation protocol, rather than establishing universal superiority over all existing MDA prediction methods.

The ablation results further clarify the contributions of the proposed components. Removing the complete hypergraph branch reduced AUC from 0.9789 to 0.8733 and AUPR from 0.9843 to 0.8896, indicating that high-order relationship modeling provides information that cannot be fully recovered by the local graph encoder. Removing the similarity hypergraph reduced AUC and AUPR to 0.9564 and 0.9673, respectively, whereas removing the interaction hypergraph resulted in values of 0.9722 and 0.9789. These results suggest that the two hypergraph branches provide complementary information, with the similarity hypergraph making the larger individual contribution under CV3. The No_Local_GCN variant remained close to the complete model, suggesting that the dynamic hypergraph modules account for most of the ranking performance, although the full model still achieved better AUC and AUPR. Furthermore, replacing the differentiable similarity hypergraph with static KNN-based hypergraphs also produced lower AUC and AUPR values than the complete model. Performance improved as the static neighborhood size increased from 5 to 15, demonstrating that fixed hypergraph performance is sensitive to the manually specified neighborhood size. Static_Int_HG produced marginally higher ACC and F1 values than DDHMDA, but these threshold-dependent differences were small relative to their standard deviations. DDHMDA retained better threshold-independent AUC and AUPR performance. Overall, the results support the use of adaptive hypergraph construction and the good performance of DDHMDA. The results of the case study indicate that DDHMDA demonstrates high reliability in the prediction of both asthma and IBD. The top 20 candidate microbes for asthma and IBD on the HMDAD dataset have a verification rate of 85% (17/20) and 100% (20/20), respectively, and the verification rate on the Disbiome dataset is also at a high level, confirming the practical value of DDHMDA in discovering potential MDAs.

Despite the promising outcomes, this study has two main limitations. First, unobserved microbe-disease pairs are not necessarily true negatives and may include undiscovered positive associations. Although multiple negative-sampling seeds and strategies were examined, false-negative label noise cannot be completely eliminated. Second, DDHMDA's

topological feature extraction mainly relies on GIP similarities, which are fundamentally derived from known interaction patterns. Consequently, for newly discovered microbes or isolated diseases entirely lacking prior known associations, their GIP representations become uninformative zero-vectors. This creates a severe cold-start problem, making it challenging for the model to accurately project these isolated nodes into the interaction hypergraph.

In conclusion, DDHMDA demonstrates superior predictive performance and provides a robust computational tool for prioritizing complex MDAs. In the future, our work will focus on the following directions. First, we plan to integrate more sophisticated negative sample selection strategies, such as clustering-based approaches [29], PU learning frameworks [25], and information-based under sampling methods [49], to screen out high-confidence negative samples and enhance the robustness of the model. Second, to overcome the cold-start problem and alleviate the reliance on topological similarities, we intend to leverage embeddings derived from pre-trained large biological models to construct richer and more intrinsic feature representations. Specifically, we will utilize genomic foundational models such as DNABERT [50] to encode microbial DNA sequences, and biomedical language models pre-trained on large-scale PubMed literature such as PubMedBERT [51,52] to extract deep textual semantics. This dual-modality pre-training strategy will effectively reduce the reliance on purely topological similarities and alleviate the cold-start problem for isolated nodes.

Author Contributions: Conceptualization, Z.L.; methodology, Z.W. and Z.L.; validation, Z.W. and Z.L.; formal analysis, Z.W.; investigation, Z.W.; resources, Z.L.; data curation, Z.W.; writing—original draft preparation, Z.W.; writing—review and editing, Z.L.; visualization, Z.W.; supervision, Z.L.; project administration, Z.L.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangdong Higher Education Young Innovative Talent Project, Grant No. 2022KQNCX204.

Data Availability Statement: The data presented in this study are openly available in HMDAD at <http://www.cuilab.cn/hmdad> (accessed on 24 March 2026) and in Disbiome at <https://disbiome.ugent.be/> (accessed on 24 March 2026). No new datasets were generated in this study; all analyses were performed using publicly available data from these two databases. The source code, configuration files, processed data, and scripts required to reproduce the five-fold cross-validation, baseline comparison, ablation experiments, and case studies are publicly available in the GitHub repository: <https://github.com/wuzhimrbeen/DDHMDA> (accessed on 28 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

To complement the main pair-level evaluation and examine generalization when no associations are available for an entire entity, Appendix A reports the supplementary cold-start experiments. CV1 evaluates held-out diseases, whereas CV2 evaluates held-out microbes. These settings are substantially more demanding than CV3 because the association-derived profiles of completely unseen entities contain little or no training signal. The corresponding results, therefore, provide a transparent assessment of the remaining cold-start limitation rather than serving as the primary comparison setting.

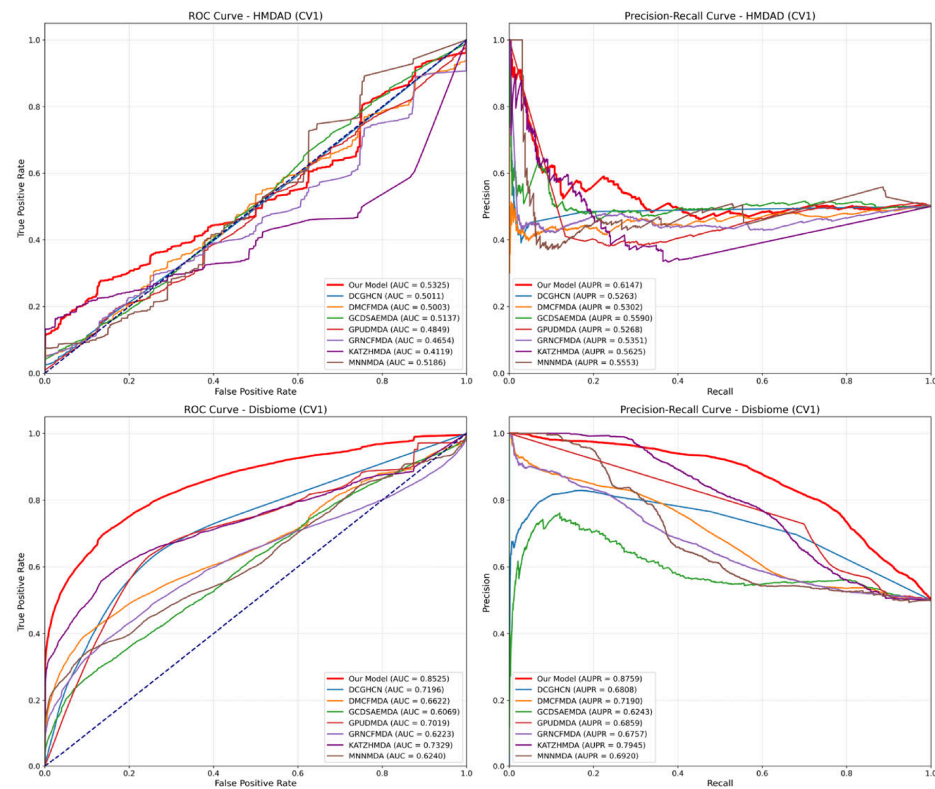


Figure A1. ROC and precision-recall curves under disease-level CV1. The upper panels show HMDAD and the lower panels show Disbiome. All associations of each held-out disease are excluded from training; ‘Our Model’ denotes DDHMDA.

Table A1. Supplementary CV1 performance on HMDAD and Disbiome.

Dataset	Method	AUC	AUPR	ACC	F1
HMDAD	DCGHCN	0.5011 ± 0.0797	0.5263 ± 0.0483	0.5000 ± 0.0000	0.6667 ± 0.0000
	DMCFMDA	0.5003 ± 0.1524	0.5302 ± 0.1096	0.4996 ± 0.1105	0.6090 ± 0.1322
	GCDSAEMDA	0.5137 ± 0.1167	0.5590 ± 0.0732	0.5468 ± 0.0708	0.6607 ± 0.0885
	GPUDMDA	0.4849 ± 0.1460	0.5268 ± 0.0798	0.5053 ± 0.1319	0.6234 ± 0.1399
	GRNCFMDA	0.4654 ± 0.1202	0.5351 ± 0.0896	0.5081 ± 0.0944	0.6296 ± 0.0934
	KATZHMDA	0.4119 ± 0.1421	0.5625 ± 0.1153	0.4909 ± 0.0250	0.6401 ± 0.0665
	MNNMDA	0.5186 ± 0.1053	0.5553 ± 0.0712	0.5564 ± 0.0631	0.6457 ± 0.0821
	DDHMDA	0.5325 ± 0.1307	0.6147 ± 0.0997	0.5179 ± 0.0839	0.6002 ± 0.1159
Disbiome	DCGHCN	0.7196 ± 0.0235	0.6808 ± 0.0363	0.6364 ± 0.0883	0.6813 ± 0.0231
	DMCFMDA	0.6622 ± 0.0324	0.7190 ± 0.0411	0.5417 ± 0.0492	0.6595 ± 0.0126
	GCDSAEMDA	0.6069 ± 0.1077	0.6243 ± 0.1000	0.5834 ± 0.0906	0.6878 ± 0.0334
	GPUDMDA	0.7019 ± 0.0538	0.6859 ± 0.0624	0.7028 ± 0.0700	0.7156 ± 0.0369
	GRNCFMDA	0.6223 ± 0.0894	0.6757 ± 0.0804	0.5476 ± 0.0802	0.6723 ± 0.0253
	KATZHMDA	0.7329 ± 0.0244	0.7945 ± 0.0225	0.7031 ± 0.0284	0.6845 ± 0.0252
	MNNMDA	0.6240 ± 0.0191	0.6920 ± 0.0186	0.5198 ± 0.0300	0.6602 ± 0.0135
	DDHMDA	0.8525 ± 0.0325	0.8759 ± 0.0301	0.7659 ± 0.0436	0.7730 ± 0.0293

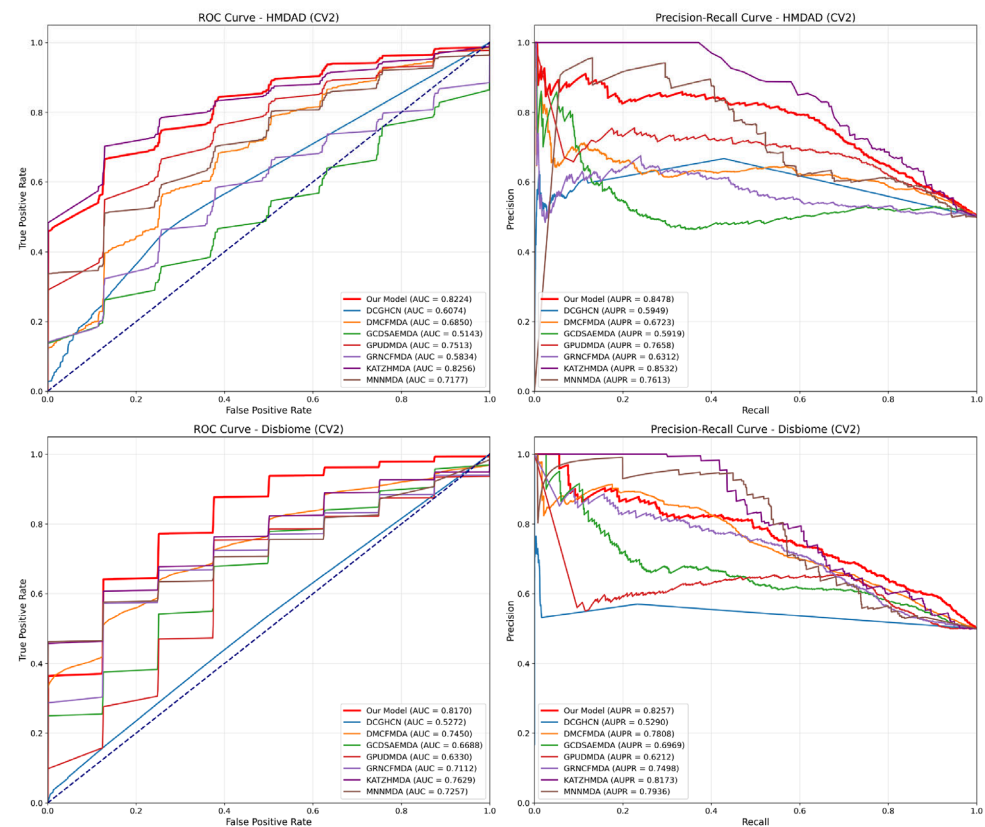


Figure A2. ROC and precision-recall curves under microbe-level CV2. The upper panels show HMDAD and the lower panels show Disbiome. All associations of each held-out microbe are excluded from training; ‘Our Model’ denotes DDHMDA.

Table A2. Supplementary CV2 performance on HMDAD and Disbiome.

Dataset	Method	AUC	AUPR	ACC	F1
HMDAD	DCGHCN	0.6074 ± 0.0910	0.5949 ± 0.0680	0.5319 ± 0.0710	0.6599 ± 0.0336
	DMCFMDA	0.6850 ± 0.1001	0.6723 ± 0.1020	0.6748 ± 0.0874	0.7205 ± 0.0467
	GCDSAEMDA	0.5143 ± 0.1805	0.5919 ± 0.1250	0.5852 ± 0.1027	0.6905 ± 0.0458
	GPUDMDA	0.7513 ± 0.0810	0.7658 ± 0.0938	0.6694 ± 0.0906	0.7014 ± 0.0559
	GRNCFMDA	0.5834 ± 0.1964	0.6312 ± 0.1446	0.6213 ± 0.1059	0.7003 ± 0.0495
	KATZHMDA	0.8256 ± 0.0286	0.8532 ± 0.0237	0.7583 ± 0.0545	0.7505 ± 0.0443
	MNNMDA	0.7177 ± 0.0740	0.7613 ± 0.0789	0.6437 ± 0.0780	0.6858 ± 0.0467
	DDHMDA	0.8224 ± 0.0365	0.8478 ± 0.0311	0.7329 ± 0.0540	0.7414 ± 0.0438
Disbiome	DCGHCN	0.5272 ± 0.0787	0.5290 ± 0.0446	0.5179 ± 0.0513	0.6693 ± 0.0137
	DMCFMDA	0.7450 ± 0.0764	0.7808 ± 0.0923	0.6816 ± 0.0489	0.7213 ± 0.0330
	GCDSAEMDA	0.6688 ± 0.0764	0.6969 ± 0.0854	0.6305 ± 0.0727	0.7026 ± 0.0340
	GPUDMDA	0.6330 ± 0.0502	0.6212 ± 0.0807	0.6504 ± 0.0596	0.6983 ± 0.0251
	GRNCFMDA	0.7112 ± 0.0524	0.7498 ± 0.0634	0.6760 ± 0.0759	0.6998 ± 0.0397
	KATZHMDA	0.7629 ± 0.0287	0.8173 ± 0.0214	0.6774 ± 0.0685	0.7089 ± 0.0404
	MNNMDA	0.7257 ± 0.0399	0.7936 ± 0.0427	0.6795 ± 0.0821	0.6861 ± 0.0392
	DDHMDA	0.8170 ± 0.0260	0.8257 ± 0.0321	0.7405 ± 0.0431	0.7722 ± 0.0296

References

- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **2012**, *486*, 207–214. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lynch, S.V.; Pedersen, O. The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **2016**, *375*, 2369–2379. [\[CrossRef\]](#) [\[PubMed\]](#)
- Aggarwal, N.; Kitano, S.; Puah, G.R.Y.; Kittelmann, S.; Hwang, I.Y.; Chang, M.W. Microbiome and human health: Current understanding, engineering, and enabling technologies. *Chem. Rev.* **2022**, *123*, 31–72. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hou, K.; Wu, Z.-X.; Chen, X.-Y.; Wang, J.-Q.; Zhang, D.; Xiao, C.; Zhu, D.; Koya, J.B.; Wei, L.; Li, J. Microbiota in health and diseases. *Signal Transduct. Target. Ther.* **2022**, *7*, 135. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gilbert, J.A.; Blaser, M.J.; Caporaso, J.G.; Jansson, J.K.; Lynch, S.V.; Knight, R. Current understanding of the human microbiome. *Nat. Med.* **2018**, *24*, 392–400. [\[CrossRef\]](#) [\[PubMed\]](#)
- Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **2014**, *16*, 276–289. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ma, W.; Zhang, L.; Zeng, P.; Huang, C.; Li, J.; Geng, B.; Yang, J.; Kong, W.; Zhou, X.; Cui, Q. An analysis of human microbe–disease associations. *Brief. Bioinform.* **2017**, *18*, 85–97. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wen, Z.; Yan, C.; Duan, G.; Li, S.; Wu, F.-X.; Wang, J. A survey on predicting microbe-disease associations: Biological data and computational methods. *Brief. Bioinform.* **2021**, *22*, bbaa157. [\[CrossRef\]](#) [\[PubMed\]](#)
- Marcos-Zambrano, L.J.; Karaduzovic-Hadziabdic, K.; Loncar Turukalo, T.; Przymus, P.; Trajkovic, V.; Aasmets, O.; Berland, M.; Gruca, A.; Hasic, J.; Hron, K. Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* **2021**, *12*, 634511. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shokri Garjan, H.; Omidi, Y.; Poursheikhali Asghari, M.; Ferdousi, R. In-silico computational approaches to study microbiota impacts on diseases and pharmacotherapy. *Gut Pathog.* **2023**, *15*, 10. [\[CrossRef\]](#) [\[PubMed\]](#)
- Garrett, W.S. Cancer and the microbiota. *Science* **2015**, *348*, 80–86. [\[CrossRef\]](#) [\[PubMed\]](#)
- Scheperjans, F.; Aho, V.; Pereira, P.A.; Koskinen, K.; Paulin, L.; Pekkonen, E.; Haapaniemi, E.; Kaakkola, S.; Eerola-Rautio, J.; Pohja, M. Gut microbiota are related to Parkinson’s disease and clinical phenotype. *Mov. Disord.* **2015**, *30*, 350–358. [\[CrossRef\]](#) [\[PubMed\]](#)
- Scher, J.U.; Sczesnak, A.; Longman, R.S.; Segata, N.; Ubeda, C.; Bielski, C.; Rostron, T.; Cerundolo, V.; Pamer, E.G.; Abramson, S.B.; et al. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2013**, *2*, e01202. [\[CrossRef\]](#) [\[PubMed\]](#)
- Visconti, A.; Le Roy, C.I.; Rosa, F.; Rossi, N.; Martin, T.C.; Mohney, R.P.; Li, W.; de Rinaldis, E.; Bell, J.T.; Venter, J.C. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **2019**, *10*, 4505. [\[CrossRef\]](#) [\[PubMed\]](#)
- Koh, A.; De Vadder, F.; Kovatcheva-Datchary, P.; Bäckhed, F. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell* **2016**, *165*, 1332–1345. [\[CrossRef\]](#) [\[PubMed\]](#)
- Matson, V.; Fessler, J.; Bao, R.; Chongsuwat, T.; Zha, Y.; Alegre, M.-L.; Luke, J.J.; Gajewski, T.F. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **2018**, *359*, 104–108. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, X.; Huang, Y.-A.; You, Z.-H.; Yan, G.-Y.; Wang, X.-S. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* **2017**, *33*, 733–739. [\[CrossRef\]](#) [\[PubMed\]](#)
- Huang, Z.-A.; Chen, X.; Zhu, Z.; Liu, H.; Yan, G.-Y.; You, Z.-H.; Wen, Z. PBHMDA: Path-based human microbe-disease association prediction. *Front. Microbiol.* **2017**, *8*, 233. [\[CrossRef\]](#) [\[PubMed\]](#)
- Guan, J.; Zhang, Z.G.; Liu, Y.; Wang, M. A novel bi-directional heterogeneous network selection method for disease and microbial association prediction. *BMC Bioinform.* **2022**, *23*, 483. [\[CrossRef\]](#) [\[PubMed\]](#)
- He, B.-S.; Peng, L.-H.; Li, Z. Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front. Microbiol.* **2018**, *9*, 2560. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yan, C.; Duan, G.; Wu, F.-X.; Pan, Y.; Wang, J. MCHMDA: Predicting microbe-disease associations based on similarities and low-rank matrix completion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 611–620. [\[CrossRef\]](#)
- Liu, H.; Bing, P.; Zhang, M.; Tian, G.; Ma, J.; Li, H.; Bao, M.; He, K.; He, J.; He, B. MNNMDA: Predicting human microbe-disease association via a method to minimize matrix nuclear norm. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 1414–1423. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, F.; Huang, Z.-A.; Chen, X.; Zhu, Z.; Wen, Z.; Zhao, J.; Yan, G.-Y. LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* **2017**, *7*, 7601. [\[CrossRef\]](#) [\[PubMed\]](#)
- Long, Y.; Luo, J.; Zhang, Y.; Xia, Y. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform.* **2021**, *22*, bbaa146. [\[CrossRef\]](#) [\[PubMed\]](#)
- Peng, L.; Huang, L.; Tian, G.; Wu, Y.; Li, G.; Cao, J.; Wang, P.; Li, Z.; Duan, L. Predicting potential microbe-disease associations with graph attention autoencoder, positive-unlabeled learning, and deep neural network. *Front. Microbiol.* **2023**, *14*, 1244527. [\[CrossRef\]](#) [\[PubMed\]](#)

26. Wang, B.; Zhao, W.; Du, X.; Zhang, J.; Zhang, C.; Wang, L.; He, Y. Predicting potential microbe-disease associations based on heterogeneous graph attention network and deep sparse autoencoder. *Eng. Appl. Artif. Intell.* **2025**, *147*, 110301. [\[CrossRef\]](#)
27. Chen, J.; Zhang, L.; Liang, Z. Utilizing dual-channel graph and hypergraph convolution network to discover microbes underlying disease traits. *J. Chem. Inf. Model.* **2025**, *65*, 5152–5162. [\[CrossRef\]](#) [\[PubMed\]](#)
28. 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. [\[CrossRef\]](#)
29. Wang, B.; Zhao, W.; Du, X.; Zhang, J.; Li, J.; Sun, H. Predicting Potential Microbe-Disease Associations Based on Heterogeneous Graph Random Attention Neural Network and Neural Collaborative Filtering. *Interdiscip. Sci. Comput. Life Sci.* **2025**, published online. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Janssens, Y.; Nielandt, J.; Bronselaer, A.; Debunne, N.; Verbeke, F.; Wynendaele, E.; Van Immerseel, F.; Vandewynckel, Y.-P.; De Tré, G.; De Spiegeleer, B. Disbiome database: Linking the microbiome to disease. *BMC Microbiol.* **2018**, *18*, 50. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Wang, J.Z.; Du, Z.; Payattakool, R.; Yu, P.S.; Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **2007**, *23*, 1274–1281. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Shi, K.; Huang, K.; Li, L.; Liu, Q.; Zhang, Y.; Zheng, H. Predicting microbe–disease association based on graph autoencoder and inductive matrix completion with multi-similarities fusion. *Front. Microbiol.* **2024**, *15*, 1438942. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Yin, W.; Wang, S.; Zhang, Y.; Qiao, S.; Wang, S.; Khan, F.; Altuki, R.; Lyu, Z. Autonomously Adjusting Multi-Relational Hypergraphs Structure for Predicting circRNA-MiRNA Associations. *IEEE J. Biomed. Health Inform.* **2025**, 1–10. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Yin, W.; Wang, S.; Zhang, Y.; Qiao, S.; Wu, W.; Li, H. Multirelational hypergraph representation learning for predicting circRNA-miRNA associations. *J. Chem. Inf. Model.* **2024**, *64*, 8349–8360. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Wilder, B.; Ewing, E.; Dilkina, B.; Tambe, M. End to end learning and optimization on graphs. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 4674–4685.
36. Papi, A.; Brightling, C.; Pedersen, S.; Reddel, H. Seminar asthma. *Lancet* **2018**, *391*, 783–800. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Arrieta, M.-C.; Stiemsma, L.T.; Dimitriu, P.A.; Thorson, L.; Russell, S.; Yurist-Doutsch, S.; Kuzeljevic, B.; Gold, M.J.; Britton, H.M.; Lefebvre, D.L. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **2015**, *7*, 307ra152. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Van Nimwegen, F.A.; Penders, J.; Stobberingh, E.E.; Postma, D.S.; Koppelman, G.H.; Kerkhof, M.; Reijmerink, N.E.; Dompeling, E.; Van Den Brandt, P.A.; Ferreira, I. Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. *J. Allergy Clin. Immunol.* **2011**, *128*, 948–955.e3. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Stiemsma, L.T.; Arrieta, M.-C.; Dimitriu, P.A.; Cheng, J.; Thorson, L.; Lefebvre, D.L.; Azad, M.B.; Subbarao, P.; Mandhane, P.; Becker, A. Shifts in *Lachnospira* and *Clostridium* sp. in the 3-month stool microbiome are associated with preschool age asthma. *Clin. Sci.* **2016**, *130*, 2199–2207. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Vael, C.; Vanheirstraeten, L.; Desager, K.N.; Goossens, H. Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* **2011**, *11*, 68. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Gao, Y.; Nanan, R.; Macia, L.; Tan, J.; Sominsky, L.; Quinn, T.P.; O’Hely, M.; Ponsonby, A.-L.; Tang, M.L.; Collier, F. The maternal gut microbiome during pregnancy and offspring allergy and asthma. *J. Allergy Clin. Immunol.* **2021**, *148*, 669–678. [\[CrossRef\]](#) [\[PubMed\]](#)
42. 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*, 509–524. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Xie, H.; Yu, S.; Tang, M.; Xun, Y.; Shen, Q.; Wu, G. Gut microbiota dysbiosis in inflammatory bowel disease: Interaction with intestinal barriers and microbiota-targeted treatment options. *Front. Cell. Infect. Microbiol.* **2025**, *15*, 1608025. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Sokol, H.; Seksik, P.; Furet, J.; Firmesse, O.; Nion-Larmurier, I.; Beaugerie, L.; Cosnes, J.; Corthier, G.; Marteau, P.; Doré, J. Low counts of *Faecalibacterium prausnitzii* in colitis microbiota. *Inflamm. Bowel Dis.* **2009**, *15*, 1183–1189. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Prosberg, M.; Bendtsen, F.; Vind, I.; Petersen, A.M.; Gluud, L.L. The association between the gut microbiota and the inflammatory bowel disease activity: A systematic review and meta-analysis. *Scand. J. Gastroenterol.* **2016**, *51*, 1407–1415. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Zhou, Y.; Zhi, F. Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: A meta-analysis. *BioMed Res. Int.* **2016**, *2016*, 5828959. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Larsen, J.M. The immune response to *Prevotella* bacteria in chronic inflammatory disease. *Immunology* **2017**, *151*, 363–374. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Bailey, J.; Vince, V.; Williams, N.; Cogan, T. *Streptococcus thermophilus* NCIMB 41856 ameliorates signs of colitis in an animal model of inflammatory bowel disease. *Benef. Microbes* **2017**, *8*, 605–614. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Qin, H. Maximal Information Coefficient-Based Undersampling Method for Highly-Imbalanced Learning. *IEEE Access* **2025**, *13*, 4126–4135. [\[CrossRef\]](#)
50. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [\[CrossRef\]](#) [\[PubMed\]](#)

51. Saqib, H.M.; Mustafa, G. UMLS-Augmented PubMedBERT for Clinical Note Diagnosis Classification: A Comparative Study. *J. Comput. Biomed. Inform.* **2025**, *10*. [[CrossRef](#)]
52. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* **2021**, *3*, 1–23. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.