IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 1 (May. – June. 2025), PP 30-37
www.iosrjournals.org

Predicting Microbe-Disease Associations Using Weighted K Nearest Known Neighbors And Multiple Kernel-Based Graph Attention Networks

Minh Tuan Lau[1], Thi Kim Anh Nguyen[1], Thi Na Nguyen[1], Quoc Phu Nguyen[1], Thi Tu Khuyen Ngo[1], Van Tinh Nguyen[1,*] And Thanh Trung Giang[1]
(School Of Information And Communications Technology, Hanoi University Of Industry, Hanoi, Vietnam.)

Abstract:
Understanding mechanisms of diseases, improving diagnosis, and identifying therapeutic targets have been indicated to rely on microbe-disease associations. However, determining these associations by traditional experiments methods is costly, laborious and tedious. Therefore, it is needed to have computational approaches for unveiling microbe-disease interactions. In this paper, we proposed new approach which combines a weighted K nearest known neighbors (WKNKN) algorithm to address the issue of sparsity of the examined datatset and multiple kernel-based graph attention networks to expose latent microbe-disease relationships. It also incorporates multi-similarity integration to improve performance of prediction. It reaches a remarkable performance demonstrating by the averaged AUC and AUPR equaling to 0.985 and 0.968, respectively, which come from results of the 10-fold cross validation experiments on the examined HMDAD dataset. These value are more forceful when comparing to some state-of-the-art approaches on same examined HMDAD dataset. Thereby, it might be recognized as a prominent tool for determining microbe-disease associations.

Key Word: *Deep learning; Graph attention networks; Multiple kernels; Predicting microbe–disease associations; WKNKN;*

Date of Submission: 22-04-2025 Date of Acceptance: 02-05-2025

I. Introduction

Microbe communities are tiny living things that can live in multicellular colonies or singlecelled organisms [1]. This varied group closely interacts with human hosts including bacteria, fungi, viruses, archaea, and protozoa, and so forth [2]. These microbial populations are found in the gastrointestinal system, skin, lungs, oral cavity and other organs [3]. Although the majority of microorganisms are helpful or beneficial, forming mutualistic associations with their hosted human bodies. However, disruptions in microbial balance have been increasingly linked to different diseases such as inflammatory bowel disease, liver disorders, diabetes and certain cancers [4]–[7]. In recent years, the critical roles of these microorganisms have been indicated in many studies, but a thorough understanding of their mechanisms in health and disease remains limited [2]. Examining microbe-disease interactions is essential not only for uncovering the complex mechanisms of disease progression but also for revealing biomarkers that could enhance diagnosis and prognosis [5]. Traditional laboratory experiments for determining associations between microbes and diseases are laborious, costly and tedious. Therefore, computational approaches leveraging machine learning, deep learning, and large-scale biological data have emerged as effective alternatives, driving significant methodological advances in this field. They can generally be categorized into the groups listed below: network-based, matrix factorization-based and neural network-based methods [8].

First of all, network-based methods are popular because of their simplicity, interpretability, and reliance on fewer parameters. However, their predictive accuracy is constrained by the availability of known associations, limiting their applicability to new diseases or microbes with no prior connections in the network. Notable methods in this category include a KATZHMDA model, presented by *Chen et al.* [9], which utilizes heterogeneous networks for identifying microbe–disease associations. Another prominent ABHMDA method fuses microbial similarity based on symptom-based disease similarity combining with Gaussian interaction profile for constructing training sample features [10]. A new method LGRSH which developed by *Lei et al.* [11] utilizes an enhanced rule-based inference and a node2vec algorithm to uncover microbe-disease interactions.

Second, matrix factorization-based methods intend to fall apart an input matrix into two lower-dimensional matrices whereas maintaining the original structure's essential properties. These approach allow for better generalization and prediction of new associations. For instance, *Peng et al.* [12] created a RNMFMDA

approach combining a matrix factorization technique of neighborhood-regularized logistic and a random walk with restart algorithm in order to forecast microbe–disease interactions. Another significant contribution is CMFHMDA method which utilizes collaborative matrix factorization for recovering matrix of microbe-disease associations [13]. Despite enabling novel association prediction, matrix factorization methods often yield suboptimal performance due to limited data representation from simplistic similarity aggregation.

Lastly, neural network-based methods have demonstrated superior performance compared to traditional approaches, offering higher prediction accuracy and the capacity to model complex relationships. One of the most notable frameworks is GATMDA, introduced by *Long et al.* [14], which predicts microbe–disease relationships using an inductive matrix completion technique and graph attention network (GAT). By incorporating attention mechanisms, GATMDA identifies the most relevant nodes and edges in the graph, thereby enhancing predictive accuracy. Another innovative model called MVGCNMDA, which was created by *Hua et al* [15]. MVGCNMDA enhances performance via multi-view graph attention. Normally, neural network-based methods excel in predictive accuracy due to their ability to model complex relationships in large datasets. Nevertheless, their reliance on latent representations reduces model interpretability.

Notably, all these groups of approaches face challenges posed by the sparsity and incompleteness of microbe–disease datasets, which hinders the robustness and generalizability of predictive models. Recently, a WKNKN algorithm and a Collaborative Filtering (CF) have been employed to solve the issue of sparsity between biological objects [16]–[18]. Additionally, as a prominent deep learning model, GAT has demonstrated its effectiveness in divergent graph-based tasks including text classification, link prediction and recommender systems. It represents microbe–disease interactions as a graph, where microbes as well as diseases are modeled as nodes while their associations form edges. By dynamically assigning attention weights to nodes and edges, GAT captures complex, non-linear relationships, enabling more accurate predictions [14]. For instance, *Wang et al.* [7] created a MKGAT framework, which integrates GAT with dual Laplacian regularized least squares for revealing associations. Additionally, the GATMDA model, introduced by *Long et al.* [14], utilizes GAT combined with inductive matrix completion for unveiling microbe-disease associations.

In this paper, to mitigate the issue of data sparsity and leverage the advancements of GATs in order to infer microbe-disease associations, we proposed a new method which combines a WKNKN algorithm and multiple kernel-based graph attention networks. The WKNKN algorithm imputes missing links by leveraging the similarity of neighboring nodes, thereby enriching the dataset with plausible associations. The integration of GAT with WKNKN creates a more robust prediction framework, allowing the model to generalize better performance which was demonstrated by AUC and AUPR values of 0.985 and 0.968, respectively. It could be recognized as an effective method for uncovering microbe-disease interactions.

II. Material And Methods

Materials

We utilized a benchmark dataset named HMDAD to assess the prediction performance. It was gathered from the Human Microbe–Disease Association Database (HMDAD, https://www.cuilab.cn/hmdad). The dataset was used in many studies including *Huang et al.*, *Wu et al.* and *Liu et al.* [19]–[21]. It contains 450 known interactions between 292 microbes and 39 diseases. In this paper, we used N_m and N_d to reflects the microbe number and disease quantity, respectively, and an adjacency matrix $A^{MD} \in R^{Nm \times Nd}$ is considered to the known microbe-disease association, in which each item $A^{MD}(i,j) \in \{0,1\}$, N_m indicates microbes' quantity while N_d represents the diseases' number. In case that a microbe has been etablished to be related to a particular disease, the $A^{MD}(i,j)$ at the respective position is set to 1, otherwise $A^{MD}(i,j) = 0$. Addionally, we used the microbe functional similarity which was calculated by *Kamneva et al.* [22] and was downloaded from the work of *Liu et al.*[21]. Additionally, we denoted the microbe functional similarity matrix as Si_m^{fun} that $Si_m^{fun}(m_i, m_j)$ represents the microbe m_i and microbe m_j similarity. In addition, it was founded on a presumption that similar diseases likely to relate with comparable genes [23], [24] and based on the HumanNet v2.0 database [25], *Liu et al.* [21] also computed the disease functional similarity matrix Si_d^{fun} that $Si_d^{fun}(d_i, d_j)$ illustrates the similarity of disease d_i and disease d_j.

Method

Our proposed method's workflow is depicted in Figure 1 and contains following stages.

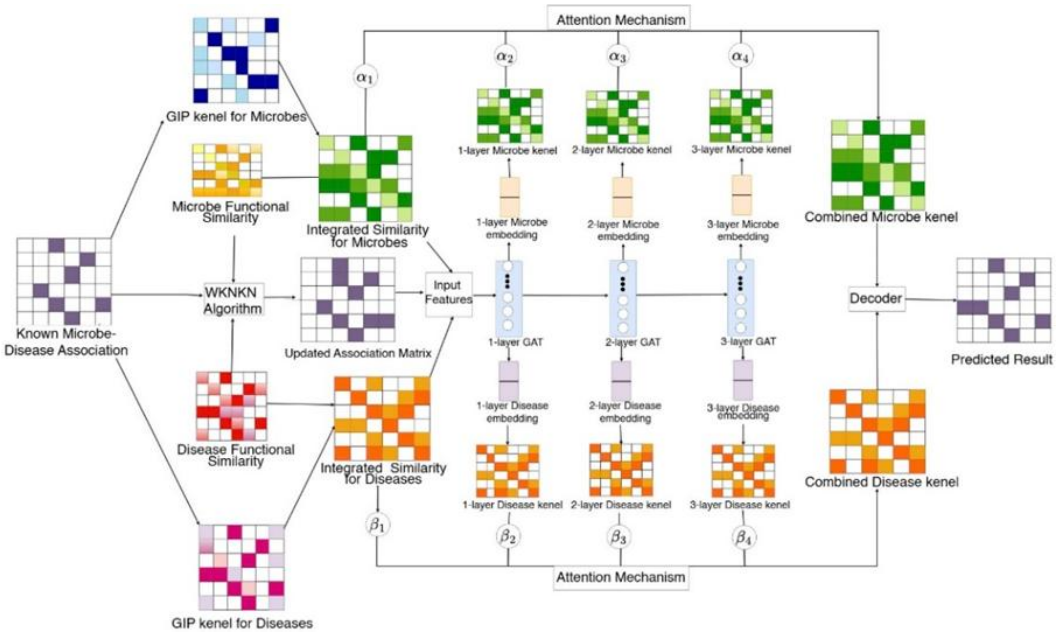

Figure 1: The workflow of proposed method

Gaussian Interaction Profile Kernel for Microbes and Diseases

In this study, apart from the microbe functional similarity as well as disease functional similarity, we also used Gaussian Interaction Profile (GIP) kernel to calculate similarity for both diseases and microbes considering its impressive performance capabilities. In the adjacent matrix A^{MD} , each i^{th} row defines the vector associated with the associations between disease and all microbes, denoted as $A^{MD}(d_i)$. In the same way, we also considered each j^{th} column as the vector associated with microbe m_j in A^{MD} , denoted as $A^{MD}(m_j)$. Similar to [21], we can calculate GIP kernel similarity for both diseases and microbes as below.

GIP kernel similarity for diseases d_i and d_j is computed as:

$GIP_{dSim}(d_i, d_j) = \exp(-\gamma_{dSim}\|A^{MD}(d_i) - A^{MD}(d_j)\|^2)$ (1)

where $-\gamma_{dSim}$ is the normalized kernel bandwidth adjustment parameter and can be computed as follow:

$\gamma_{dSim} = \gamma'_{dSim} / \frac{1}{N_d}\Sigma_{i=1}^{N_d}\left\|A^{MD}(d_i)\right\|^2$ (2)

where N_d indicates the quantity of diseases, γ'_{dSim} is the original bandwidth. The value of γ'_{dSim} was fixed to 1 as in the study of *Zhang et al.* [26]. Similarly, with the microbes m_i and m_j, GIP kernel similarity was calculated as:

$GIP_{mSim}(m_i, m_j) = \exp(-\gamma_{mSim}\|A^{MD}(m_i) - A^{MD}(m_j)\|^2)$ (3)

where $-\gamma_{mSim}$ is the normalized kernel bandwidth adjustment parameter and can be computed as:

$\gamma_{mSim} = \gamma'_{mSim} / \frac{1}{N_m}\Sigma_{i=1}^{N_m}\left\|A^{MD}(m_i)\right\|^2$ (4)

where N_m indicates the quantity of microbes, γ'_{mSim} is the original bandwidth. The value of γ'_{mSim} was fixed to 1 as in the the study of *Zhang et al.* [26].

Calculating Integrated Similarity for Microbes and diseases

To explore new relationships more effectively, we integrated Microbe functional similarity with Microbe GIP kernel similarity for attaining the Integrated Similarity matrix for Microbes (ISM) as:

$$\mathrm{ISM}(m_i, m_j) = \begin{cases} (\mathrm{Si}_{\mathrm{m}}^{\mathrm{fun}}(mi, mj) + GIPmSim(mi, mj))/2 \; if \; \mathrm{Si}_{\mathrm{m}}^{\mathrm{fun}}(\mathrm{mi}, \mathrm{mj}) \neq 0 \\ GIP_{mSim}(m_i, m_j) \quad \text{otherwise} \end{cases} \quad (5)$$

Similarly, disease functional similarity was also integrated with disease GIP kernel similarity for attaining the Integrated Similarity matrix for Diseases (ISD) as:

$$\mathrm{ISD}(d_i, d_j) = \begin{cases} (\mathrm{Si}_{\mathrm{d}}^{\mathrm{fun}}(d_i, \mathrm{d_j}) + GIPdSim(d_i, d_j))/2 \; if \; \mathrm{Si}_{\mathrm{d}}^{\mathrm{fun}}(d_i, d_j) \neq 0 \\ GIP_{dSim}(d_i, d_j) \quad \text{otherwise} \end{cases} \quad (6)$$

Improving Microbe-Disease Association Matrix Using a Weight K-Nearest Known Neighbour Algorithm

As be mentioned, the validated microbe-disease association matrix is sparse. To lessen the sparsity in association matrix, we pre-processed it using a WKNKN algorithm inspiring by the success of previous studies [16], [27], [28]. WKNKN follows the same principle of K nearest neighbors (KNN), but adds a "weighting"

mechanism on how to process the missing data cases by borrowing from nearest neighbors that can potentially be truthful associations. Firstly, for each microbe, we utilized the functional similarity matrix $Si_m^{fun}(m_i, m_j)$ to figure out known microbes adjacent to it (k), and used the scores similarity of them to deduce the interaction possibility potentially of m_i, based on [28]:

$$A_m^{MD}(m_i, :) = \frac{1}{Z_m}\Sigma_{n_m=1}^{k} W_{n_m} A_m^{MD}(m_{n_m}, :)$$

$$W_{n_m} = \alpha^{n_m-1} Si_m^{fun}(m_i, m_{n_m})$$

in which $\alpha \leq 1, Z_m = \Sigma_{n_m=1}^{k} Si_m^{fun}(m_i, m_{n_m})$ (7)

that m_{n_1} to m_{n_k} are the k nearest known neighbors of m_i, sorted in decreasing order, W_{n_m} is the weight coefficient.

Secondly, we continued to use the semantic similarity matrix $Si_d^{fun}(d_i, d_j)$ to infer the probability spectrum of disease's interaction, below:

$$A_d^{MD}(:, d_j) = \frac{1}{Z_d}\Sigma_{n_d=1}^{k} W_{n_d} A_d^{MD}(:, d_{n_d})$$

$$W_{n_d} = \alpha^{n_d-1} Si_d^{fun}(d_{n_d}, d_j)$$

in which $\alpha \leq 1, Z_d = \Sigma_{n_d=1}^{k} Si_d^{fun}(d_{n_d}, d_j)$ (8)

where d_{n_1} to d_{n_k} are the k nearest known neighbors of d_i, sorted in decreasing order, W_{n_d} is the weight coefficient.

Finally, we used the average of A_m^{MD} and A_d^{MD}, and replace blanks with the corresponding values, in two equations below:

$$T = (A_m^{MD} + A_d^{MD})/2 \quad (9)$$

$$A_{i,j}^{MD} = \begin{cases} T_{i,j} & if\ A_{i,j}^{MD} = 0 \\ A_{i,j}^{MD} & if\ A_{i,j}^{MD} \neq 0 \end{cases} \quad (10)$$

where we defined A^{MD_new} as a matrix processed by the WKNKN algorithm.

The Heterogeneous Network and the GAT Architecture

Inspired by the work of *Wang and Chen* [7], we built the heterogeneous network as:

$$HN = \begin{pmatrix} ISM & A^{MD\_new} \\ (A^{MD\_new})^T & ISD \end{pmatrix} \quad (11)$$

where HN is the heterogeneous network, ISM is the matrix of integrated similarity for microbes, ISD is the matrix of integrated similarity for diseases, and A^{MD_new} is the adjacency matrix obtaining from WKNKN while its transpose matrix is $(A^{MD\_new})^T$.

GAT is known as a network that works with graph data [29], in this work, it is deployed to acquire the features of microbes as well as diseases. In particular, given the HN matrix above, GAT can be formulated as:

$$H^{(l)} = f(H^{(l-1)}, HN) = \sigma(GAT(H^{(l-1)}, HN)) \quad (12)$$

where $H^{(l)}$ represents the *l*-layer embedding of nodes, with *l* =1,...,L, $\sigma(\ )$ means a non-linear activation function (ReLU). A single graph attention layer is indicated by GAT and the full L-layer structure is piled up by numerous GATs. The node feature set h = $\{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}\}$ is used as the initial input, $\overrightarrow{h_i} \in R^F$ where n is the nodes'quantity and F reflects the quantity of features per node. A new node feature set is produced by this layer $h' = \{\overrightarrow{h_1'}, \overrightarrow{h_2'}, \ldots, \overrightarrow{h_n'}\}$, $\overrightarrow{h_i'} \in R^{F'}$.These features are transformed into a upper-level by utilizing a learnable linear transformation, where the weight matrix $W \in R^{F' \times F}$ is applied to separate node. Following this, the attention factors were measured as:

$$e_{ij} = a(W\overrightarrow{h_i}, W\overrightarrow{h_j}) \quad (13)$$

We applied the softmax function to normalize the attention factors to attain the coefficients as:

$$\alpha_{ij} = softmax(e_{ij}) = exp(e_{ij})/\Sigma_{k\epsilon N_i} exp(e_{ik}) \quad (14)$$

The attention mechanism coefficients are calculated by substituting (13) into (14) as follows:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(\text{LeakyReLU}(\vec{a}^T, [W\overrightarrow{h_i} \,||\, W\overrightarrow{h_j}]))}{\Sigma_{k\epsilon N_i} exp(\text{LeakyReLU}(\vec{a}^T, [W\overrightarrow{h_i} \,||\, W\overrightarrow{h_j}]))} \quad (15)$$

Where, α is the attention coefficient, $\vec{a} \in R^{2F'}$ refers to the weight vector that is parameterized, LeakyReLu indicates the activation function, || denotes the connection operation, and N_i indicates the group of neighbors associated with node i. Following the calculation of the normalized attention coefficients, the each node's final output features are obtained:

$$\overrightarrow{h_i'} = \alpha_{ii} W\overrightarrow{h_i} + \Sigma_{j\,\epsilon N_i} \alpha_{ij}\, W\overrightarrow{h_j} \quad (16)$$

In this study, we generated the initial embedding $H^{(0)}$ as:

$$H^{(0)} = \begin{pmatrix} 0 & A^{MD\_new} \\ (A^{MD\_new})^T & 0 \end{pmatrix} \quad (17)$$

Multi-kernel integration

The multiple-layer GAT model can generate multi-embeddings, each representing distinct graph structures. Different structural information is reflected by these embeddings, enabling the resulting kernels to represent node similarities from various perspectives. The kernel sets for the microbe space $S^M = \{S_m, K_{h1}^m, \ldots, K_{hL}^m\}$ and the disease space $S^D = \{S_d, K_{h1}^d, \ldots, K_{hL}^d\}$ can be constructed by incorporating the existing similarity matrix. Where, K_{hl}^m , K_{hl}^d denote the kernel matrices corresponding to the microbe and disease embeddings for each layer. We integrated multi-kernel as:

$K_m = \sum_{i=1}^{L+1} AM_i^m S_i^m$ (18)

$K_d = \sum_{i=1}^{L+1} AD_i^d S_i^d$ (19)

where, S_i^m and S_i^d refer to the i-th kernel in the microbe and disease kernels. AM_i and AD_i represent the attention coefficients associated with each kernel, while L means the layers' total number.

Decoder for Microbe-Disease Association Prediction

To obtain associations between microbes and diseases, we utilized the Dual Laplacian regularized least squares with two microbe and disease feature spaces combined kernel matrices [7]. In this study, the loss function is as:

$\min J = \|K_m\alpha_m + (K_d\alpha_d)^T - 2A^{MD\_new}_{train}\|_F^2 + \varphi(\|\alpha_m\|_F^2 + \|\alpha_d\|_F^2)$ (20)

where, $\|\ \|_F$ represents the Frobenius norm, $A^{MD\_new}_{train} \epsilon\ R^{Nm \times Nd}$ is the adjacency matrix, $\alpha_m, \alpha_d{}^T \epsilon\ R^{Nm \times Nd}$ is the trainable matrix, $K_m \epsilon\ R^{Nm \times Nm}$ and $K_d \epsilon\ R^{Nd \times Nd}$ form the combined kernel sets in the two feature spaces, φ is a decay coefficient that adjusts the regularization term's weight. Thus, the result of microbe-disease associations predicted across the two feature spaces is as below:

$F^* = K_m\alpha_m + (K_d\alpha_d)^T/2$ (21)

III. Experimental Results

Parameter settings

In this paper, graph attention networks (GATs) are employed to derive features related to microbe-disease associations. A multi-layer GAT model is specified for the determination of embeddings at multiple levels. Further, graph embeddings and fused similarities are used in combining multiple kernel matrices. Specifically, the number of GAT layers L is defined as 3. The embedding dimensions (k1, k2, k3) for each layer are configured as 256, 64 and 32. The learning rate lr is assigned a value of 0.001, while the regularization parameters λ_1 and λ_2 are set to 2^{-3} and 2^{-4}, and the values for γ_h_1, γ_h_2, and γ_h_3 are specified as 2^{-4}, 2^{-3}, and 2^{-5}. For parameter of k nearest known neighbors, the optimized k value is set to 5 after performing repeatedly experiments.

Performance Evaluation

We conducted 10-fold cross-validation experiments in order to measure the predictive performance by computing the our method AUC and AUPR values on the HMDAD dataset as previously stated. We calculated the TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) for the obtained results in each experimental running time. Specifically, TP and TN demonstrate the nicely predicted positive and negative samples, whereas FP and FN indicate the misclassified positive and negative samples. Our method performance was assessed through plotting the Receiver Operating Characteristics (ROC) [30] and Precision-Recall (P-R) [31]curves. The ROC curves were plotted on TPR, FPR while P-R curves utilizing Precision and Recall. They were computed as:

$FPR = FP/(TN+FP)$ (22)

$TPR = TP/(TP+FN)$ (23)

$Recall = TP/(TP+FN)$ (24)

$Precision = TP/(TP+FP)$ (25)

Figure 2 shows our method ROC curves and AUCs whereas Figure 3 illustrates P-R curves and AUPRs when performing 10-fold cross validation experiments and the number of k nearest neighbors is set to 5. Different AUC and AUPR values when k is changed from 2 to 7 are demonstrated in Table 1

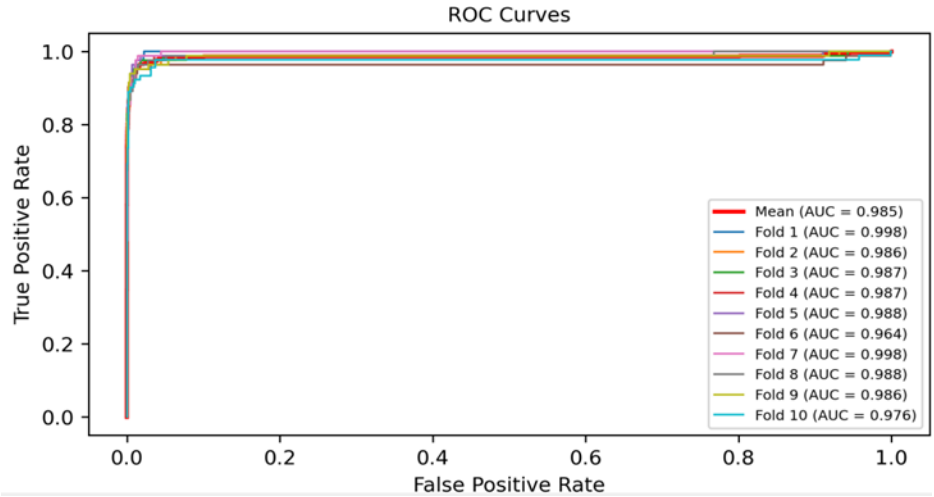

Figure 2: Our method's ROC curves and AUC values in 10 running time of 10-fold cross experiments when k=5

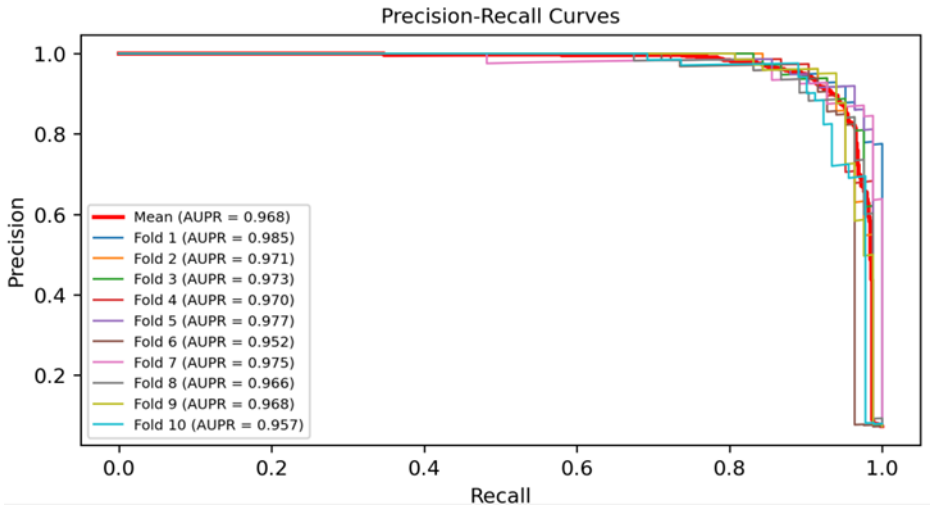

Figure 3: Our method's P-R curves and AUPR values in 10 running time of 10-fold cross experiments when k=5

Table 1: Our method's AUC and AUPR values when k is changed from 2 to 7

Value of k	AUC	AUPR
2	0.978	0.928
3	0.977	0.932
4	0.981	0.951
5	0.985	0.968
6	0.983	0.954
7	0.973	0.941

Ablation studies

For the purpose of understanding the impacts of the integration of multiple similarities and also utilizing the WKNKN algorithm, we have conducted some ablation studies and summarized the results as in Table 2 below.

Table 2: Our method's AUC and AUPR values in some ablation case studies.

Ablation case study	AUC	AUPR
No WKNKN, only functional similarity of microbes and diseases	0.949	0.784
No WKNKN, integrating similarity	0.969	0.898
WKNKN, k=5, only functional similarity of microbes and diseases	0.966	0.894
WKNKN, k=5, integrating similarity (proposed method)	**0.985**	**0.968**

Comparison with other methods

For the purpose of illustrating that our method's prediction performance is superior to some recent approaches, we utilized the results reported by the authors in their respective papers [14], [15], [32]. The experiments in these studies are performed on the same HMDAD dataset as our method. Both our method's AUC and AUPR values are higher than all of the other related approaches as demonstrated in Table 3. It highlighted that it is potential an valuable tool for revealing microbe-disease associations

Table 3: Our method's prediction performance and other methods' prediction performance on the same HMDAD dataset under 10-fold cross validation experiments.

Method	AUC	AUPR
GATMDA[14]	0.9398	0.9364
MVGCNMDA [15]	0.9196	0.9327
MVFA [32]	0.9718	0.8864
Our proposed method	0.985	0.968

Checking case study

To increase our method's reliability in revealing microbe-disease interactions, Type 2 diabetes was chosen as a case study due to its relevance and prevalence. As can be known, with a rising incidence worldwide, Type 2 diabetes is characterized by insulin resistance as well as elevated blood sugar levels [33]. Understanding the associations between Type 2 Diabetes and microbial interactions is essential for advancing treatment strategies and improving patient care. Thus, for increasing our proposed method's predictive reliability, Type 2 Diabetes was selected as a representative case study. As demonstrated in Table no 3, among the top 10 microbes predicted to be associated with Type 2 Diabetes using this approach, 8 associations have been validated through the HMDAD database.

Table no 4: The top 10 predicted Type 2 diabetes-associated microbes.

Rank	Microbe ID	Microbe name	Evidence
1	22	Bacilli	HMDAD
2	105	Clostridium	HMDAD
3	61	Betaproteobacteria	HMDAD
4	60	Bacteroidetes	HMDAD
5	230	Proteobacteria	HMDAD
6	153	Faecalibacterium prausnitzii	HMDAD
7	224	Prevotella	Not conformed
8	154	Firmicutes	HMDAD
9	108	Clostridium difficile	Not confirmed
10	183	Lactobacillus	HMDAD

IV. Conclusion And Discussions

Identifying the potential microbe-disease relationships not only facilitates diseases' diagnosis, prognosis and treatment but also contributes to microbe-targeted therapies in precision medicine. However, determining these associations is often expensive, tedious and laborious. As a result, it is imperative to advance computational methods to improve the performance of current models. In this work, we introduced a new computational approach aimed at unveiling the associations between microbes and diseases. This approach offers several key contributions. Firstly, using the WKNKN to handle data sparsity and integrate multiple information sources improves prediction performance. Secondly, processing complex graph data with Graph Attention Networks (GAT) improves the model's capacity to identify complex relationships and rank important aspects, allowing for more precise and effective analysis. Finally, our approach can perform better than other state-of-the-art techniques for forecasting possible microbe-disease associations. Even with our model's high predictive performance, there are still some drawbacks. The intricacy and unpredictability of microbe-disease associations, as well as the scarcity and lack of validation of pertinent microbial data, provide difficulties. To improve the model's ability to capture and interpret these associations more successfully, future research should place a high priority on incorporating complicated network embedding approaches, such as knowledge graphs.

References

[1] R. Ley, "The Human Microbiome: There Is Much Left To Do," Nature, Vol. 606, No. 7914. P. 435, 2022. Doi: 10.1038/D41586-022-01610-5.

[2] Z. Wen, C. Yan, G. Duan, S. Li, F. X. Wu, And J. Wang, "A Survey On Predicting Microbe-Disease Associations: Biological Data And Computational Methods," Brief. Bioinform., Vol. 22, No. 3, Pp. 1–20, (2021), Doi: 10.1093/Bib/Bbaa157.

[3] E. Holmes, A. Wijeyesekera, S. D. Taylor-Robinson, And J. K. Nicholson, "The Promise Of Metabolic Phenotyping In Gastroenterology And Hepatology." Pp. 458–471, 2015.

[4] L. Peng Et Al., "Analysis Of Ct Scan Images For Covid-19 Pneumonia Based On A Deep Ensemble Framework With Densenet,

Swin Transformer, And Regnet," Front. Microbiol., Vol. 13, No. September, Pp. 1–14, (2022), Doi: 10.3389/Fmicb.2022.995323.
[5] A. Roque Et Al., "Dietary Patterns Drive Loss Of Fiber-Foraging Species In The Celiac Disease Patients Gut Microbiota Compared To First-Degree Relatives," Gut Pathog., Vol. 16, No. 1, (2024), Doi: 10.1186/S13099-024-00643-7.
[6] L. Wen Et Al., "Innate Immunity And Intestinal Microbiota In The Development Of Type 1 Diabetes," Nature, Vol. 455, No. 7216, Pp. 1109–1113, (2008), Doi: 10.1038/Nature07336.
[7] W. Wang And H. Chen, "Predicting Mirna-Disease Associations Based On Graph Attention Networks And Dual Laplacian Regularized Least Squares," Brief. Bioinform., Vol. 23, No. 5, Pp. 1–13, (2022), Doi: 10.1093/Bib/Bbac292.
[8] H. Zhu, H. Hao, And L. Yu, "Identifying Disease-Related Microbes Based On Multi-Scale Variational Graph Autoencoder Embedding Wasserstein Distance," Bmc Biol., Vol. 21, No. 1, Pp. 1–15, (2023), Doi: 10.1186/S12915-023-01796-8.
[9] X. Chen, Y. A. Huang, Z. H. You, G. Y. Yan, And X. S. Wang, "A Novel Approach Based On Katz Measure To Predict Associations Of Human Microbiota With Non-Infectious Diseases," Bioinformatics, Vol. 33, No. 5, Pp. 733–739, (2017), Doi: 10.1093/Bioinformatics/Btw715.
[10] L. H. Peng, J. Yin, L. Zhou, M. X. Liu, And Y. Zhao, "Human Microbe-Disease Association Prediction Based On Adaptive Boosting," Front. Microbiol., Vol. 9, No. Oct, Pp. 1–9, (2018), Doi: 10.3389/Fmicb.2018.02440.
[11] X. Lei And Y. Wang, "Predicting Microbe-Disease Association By Learning Graph Representations And Rule-Based Inference On The Heterogeneous Network," Front. Microbiol., Vol. 11, No. April, Pp. 1–10, (2020), Doi: 10.3389/Fmicb.2020.00579.
[12] L. Peng, L. Shen, L. Liao, G. Liu, And L. Zhou, "Rnmfmda: A Microbe-Disease Association Identification Method Based On Reliable Negative Sample Selection And Logistic Matrix Factorization With Neighborhood Regularization," Front. Microbiol., Vol. 11, No. October, Pp. 1–12, (2020), Doi: 10.3389/Fmicb.2020.592430.
[13] J. Chen, R. Tao, Y. Qiu, And Q. Yuan, "Cmfhmda: A Prediction Framework For Human Disease-Microbe Associations Based On Cross-Domain Matrix Factorization," Brief. Bioinform., Vol. 25, No. 6, (2024), Doi: 10.1093/Bib/Bbae481.
[14] Y. Long, J. Luo, Y. Zhang, And Y. Xia, "Predicting Human Microbe-Disease Associations Via Graph Attention Networks With Inductive Matrix Completion," Brief. Bioinform., Vol. 22, No. 3, Pp. 1–13, (2021), Doi: 10.1093/Bib/Bbaa146.
[15] M. Hua, S. Yu, T. Liu, X. Yang, And H. Wang, "Mvgcnmda: Multi-View Graph Augmentation Convolutional Network For Uncovering Disease-Related Microbes," Interdiscip. Sci. – Comput. Life Sci., Vol. 14, No. 3, Pp. 669–682, (2022), Doi: 10.1007/S12539-022-00514-2.
[16] A. Ezzat, P. Zhao, M. Wu, X. L. Li, And C. K. Kwoh, "Drug-Target Interaction Prediction With Graph Regularized Matrix Factorization," Ieee/Acm Trans. Comput. Biol. Bioinforma., Vol. 14, No. 3, Pp. 646–656, (2017), Doi: 10.1109/Tcbb.2016.2530062.
[17] V. T. Nguyen, T. T. K. Le, K. Than, And D. H. Tran, "Predicting Mirna–Disease Associations Using Improved Random Walk With Restart And Integrating Multiple Similarities," Sci. Rep., Vol. 11, No. 1, Pp. 1–16, (2021).
[18] J. Yu, Z. Xuan, X. Feng, Q. Zou, And L. Wang, "A Novel Collaborative Filtering Model For Lncrna-Disease Association Prediction Based On The Naïve Bayesian Classifier," Bmc Bioinformatics, Vol. 20, No. 1, Pp. 1–13, (2019).
[19] J. Y. Shi, H. Huang, Y. N. Zhang, J. B. Cao, And S. M. Yiu, "Bmcmda: A Novel Model For Predicting Human Microbe-Disease Associations Via Binary Matrix Completion," Bmc Bioinformatics, Vol. 19, No. Suppl 9, (2018), Doi: 10.1186/S12859-018-2274-3.
[20] C. Wu Et Al., "Gcnpmda: Human Microbe-Disease Association Prediction By Hierarchical Graph Convolutional Network With Layer Attention," Biomed. Signal Process. Control, Vol. 100, No. January 2024, (2025), Doi: 10.1016/J.Bspc.2024.107004.
[21] H. Liu Et Al., "Mnnmda: Predicting Human Microbe-Disease Association Via A Method To Minimize Matrix Nuclear Norm," Comput. Struct. Biotechnol. J., Vol. 21, Pp. 1414–1423, (2023), Doi: 10.1016/J.Csbj.2022.12.053.
[22] O. K. Kamneva, "Genome Composition And Phylogeny Of Microbes Predict Their Co-Occurrence In The Environment," Plos Comput. Biol., Vol. 13, No. 2, Pp. 1–20, (2017), Doi: 10.1371/Journal.Pcbi.1005366.
[23] H. Wei, Y. Xu, And B. Liu, "Icircda-Ltr: Identification Of Circrna-Disease Associations Based On Learning To Rank," Bioinformatics, Vol. 37, No. 19, Pp. 3302–3310, (2021), Doi: 10.1093/Bioinformatics/Btab334.
[24] J. Xu And Y. Li, "Discovering Disease-Genes By Topological Features In Human Protein-Protein Interaction Network," Bioinformatics, Vol. 22, No. 22, Pp. 2800–2805, (2006), Doi: 10.1093/Bioinformatics/Btl467.
[25] S. Hwang Et Al., "Humannet V2: Human Gene Networks For Disease Research," Nucleic Acids Res., Vol. 47, No. D1, Pp. D573–D580, (2019), Doi: 10.1093/Nar/Gky1126.
[26] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, And X. Li, "Predicting Potential Drug-Drug Interactions By Integrating Chemical, Biological, Phenotypic And Network Data," Bmc Bioinformatics, Vol. 18, No. 1, Pp. 1–12, (2017), Doi: 10.1186/S12859-016-1415-9.
[27] G. Li, J. Luo, Q. Xiao, C. Liang, And P. Ding, "Predicting Microrna-Disease Associations Using Label Propagation Based On Linear Neighborhood Similarity," J. Biomed. Inform., Vol. 82, No. May, Pp. 169–177, (2018), Doi: 10.1016/J.Jbi.2018.05.005.
[28] H. Wen, X. Zhong, L. Lin, And L. Chen, "Ans-Scmc: A Matrix Completion Method Based On Adaptive Neighbourhood Similarity And Sparse Constraints For Predicting Microbe-Disease Associations," J. Cell. Mol. Med., Vol. 28, No. 18, Pp. 1–14, (2024), Doi: 10.1111/Jcmm.70071.
[29] P. Velicković, G. Cucurull, A. Casanova, A. Romero, P. Liò, And Y. Bengio, "Graph Attention Networks," Arxiv, Pp. 1–12, (2017).
[30] H.-T. K., "Receiver Operating Characteristic (Roc) Curve Analysis For Medical Diagnostic Test Evaluation," Casp. J Intern Med 2013;, Vol. 4(2), Pp. 627–635, (2013).
[31] T. Saito And M. Rehmsmeier, "The Precision-Recall Plot Is More Informative Than The Roc Plot When Evaluating Binary Classifiers On Imbalanced Datasets," Plos One, Vol. 10, No. 3, P. E0118432., (2015), Doi: 10.1371/Journal.Pone.0118432.
[32] W. Peng, M. Liu, W. Dai, T. Chen, Y. Fu, And Y. Pan, "Multi-View Feature Aggregation For Predicting Microbe-Disease Association," Ieee/Acm Trans. Comput. Biol. Bioinforma., Vol. 20, No. 5, Pp. 2748–2758, (2023), Doi: 10.1109/Tcbb.2021.3132611.
[33] S. Chatterjee, K. Khunti, And M. J. Davies, "Type 2 Diabetes," Lancet, Vol. 389, No. 10085, Pp. 2239–2251, (2017), Doi: 10.1016/S0140-6736(17)30058-2.