Neighborhood topology-aware knowledge graph learning and microbial preference inferring for drug-microbe association prediction

Jing Gu,[†] Tiangang Zhang,[‡] Yihang Gao,[†] Sentao Chen,[¶] Yuxin Zhang,[¶] Hui Cui,[§] and Ping Xuan[*,¶]

[†]*School of Computer Science and Technology, Heilongjiang University, Harbin, 150080, China*

[‡]*School of Mathematical Science, Heilongjiang University, Harbin, 150080, China*

[¶]*Department of Computer Science and Technology, Shantou University, Shantou, 515063, China*

[§]*Department of Computer Science and Information Technology, La Trobe University, Melbourne, 3083, Australia*

E-mail: pxuan@stu.edu.cn

Abstract

The human microbiota may influence the effectiveness of drug therapy by activating or inactivating the pharmacological properties of drugs. Computational methods have demonstrated their ability in screening the reliable microbe-drug associations and uncovering the mechanism that drugs exert their functions. However, the previous prediction methods failed to completely exploit the neighborhood topologies of microbe and drug entities, and the diverse correlations between the microbe-drug entity

pair and the other entities. In addition, they ignored the case that a microbe prefers to associate with its own specific drugs. A novel prediction method, PCMDA, was proposed by learning the neighborhood topologies of entities, inferring the association preferences, and integrating the features of each entity pair based on multiple biological premises. First, a knowledge graph consisting of microbe, disease, and drug entities is established to help the subsequent integration of the topological structure of entities and the similarity, interaction, and association relationship between any two entities. We generate various topological embeddings for each microbe (or drug) entity through random walks with neighborhood-restart on the microbe-disease-drug knowledge graph. Distance-level attention is designed to adaptively fuse neighborhood topologies covering multiple ranges. Second, the topological embeddings of entities imply the latent topological relationships between entities, while the relational embeddings of entities derive from the semantics of connections among the entities. The topological structure and relational semantics of entities are fused by a designed knowledge graph learning module based on multi-layer perceptron networks. Third, considering the preference that each microbe tends to specially associate with a group of drugs, an information-level attention is designed to integrate the dependency between microbial preference and the candidate drug. Finally, a dual-gated network is established to encode the features of a microbe-drug entity pair from multiple biological perspectives. The comparative experiments with seven state-of-the-art methods demonstrate PCMDA's superior performance for microbe-drug association prediction. The case studies on three drugs and the recall rate evaluation for the top-ranked candidates indicate PCMDA has the capability in discovering reliable the candidate microbes associated with a drug.

Introduction

The human microbiome consists of bacteria, viruses, and other microbes that inhabit different parts of the human body. Extensive biological studies have shown that these organisms have a significant impact on human health. For instance, SARS-CoV-2 can cause pulmonary

inflammation and various acute sequelae.[1] Additionally, the gut microbiota can influence brain function through the gut–brain axis.[2] There is also a close connection between microbes and drugs; for example, microbes can activate drugs through chemical reactions and have become a focal point of precision pharmacology.[3] Therefore, it is crucial to identify drug-related candidate microbes for new drug research and therapy development.

With the advancement of big data and cloud computing, computational methods are increasingly being utilized to identify reliable candidate microbes for biologists to test in wet lab experiments. Long *et al.* constructed a diverse network of microbe–disease–drug interactions and predicted microbe–drug associations using a hierarchical attention mechanism.[4] Huang *et al.* employed a graph autoencoder with similarity-based features as inputs to learn the distribution of microbe–drug associations.[5] Xuan *et al.* modeled the diverse information of microbe and drug nodes and inferred their associations using diverse graph neural networks.[6] However, these methods have certain limitations. For instance, they often use similar methods to learn representations of microbes and drugs, overlooking the diversity between them. Additionally, essential information, such as the biological premises of microbe–drug associations, has not been adequately utilized, which could enhance the prediction of candidate microbes for drugs.

We propose a method called PCMDA for predicting associations between microbes and drugs. The goal is to learn entity representations that are aware of neighborhood topology from a microbe–disease–drug knowledge graph. We also aim to infer the preference features of microbes for potential drugs and encode biological correlations across multiple biological premises. The main contributions of our work are summarized below:

Firstly, a knowledge graph was established and it included the microbe entities, the disease ones, and the drug ones. It also contains the biological relationships among these entities, such as the similarity relationships, the association ones, and the interaction ones. The k-distance neighborhood of a microbe, disease, or drug entity revealed the distance among the entities, and the designed strategy by random walks with neighborhood restart on

the knowledge graph was helpful for demonstrating the multi-scale neighborhood topologies.

Secondly, since the multi-scale topologies have different contributions to the representation learning of microbe and drug entities, the distance-level attention was proposed to adaptively integrate the diverse semantics within these topological structures. A knowledge graph learning strategy based on multi-layer perceptron networks was presented to encode the topological structures of head and tail entities from both the channel and spatial perspectives, along with the multiple types of relationships between two entities.

Thirdly, a target microbe tends to be associated with a group of its specific drugs, which reflects its association preference. An information-level attention was proposed to fuse the various information of drugs which includes the similarity information, the interaction one, and the entity one. The fused information was beneficial for further inferring the tendency that a microbe with association preference is related to a candidate drug.

Finally, if they have the interaction, association, similarity relationships with more common microbes and drugs, a microbe and a drug are more likely to be associated with each other. A dual-gated network was constructed to encode the denpendencies among multiple biological correlations and disginguish the importance of channel semantic and that of spatial features. The comparion with advanced compared methods, the ablation experiments, and the case studies demonstrated the superior performance of our model and its ability in discovering the potential drug-microbe association candidates.

Materials

We have incorporated various biological datasets to create the Integrated Microbe–Drug Association Dataset (IMDAD). A total of 2,268 microbe–drug associations were obtained from MDAD,[7] aBiofilm,[8] DrugVirus,[9] and published literature. These associations encompass 1,209 drugs, 172 microbes, and 154 diseases. Among these, 452 microbe–drug associations were extracted from published literature (details in Supplementary File SF1)

using text mining tools. Additionally, we gathered 435 microbe–disease associations from gutMDisorder,[10] HMDAD[11] and Peryton,[12] as well as 700 drug–disease associations from CTD.[13] Furthermore, we retrieved 10,783 drug–drug and 109 microbe–microbe interactions from DrugBank[14] and MIND (https://visant-new.bu.edu/mind/). SMILES strings were obtained from DrugBank and aBiofilm, and gene sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). Based on the chemical structures of drugs, we utilized the SIMCOMP[15] tool to assess the degree of similarity between different drugs. We calculate the semantic similarity between diseases based on their directed acyclic graphs, which are constructed from the MeSH tree structures of the diseases. The consistency of microbial gene sequences reflects their similarity, with higher consistency indicating a higher probability of homology among the microbes. We integrate the cosine similarity[6] and the identity of gene sequences[16] as the microbial similarity $K^{micr} \in R^{N^{micr} \times N^{micr}}$,

$$K_{ij}^{micr} = \begin{cases} K_{ij}^{ani}, & K_{ij}^{ani} \neq 0 \\ K_{ij}^{cos}, & K_{ij}^{ani} = 0 \end{cases}, \tag{1}$$

where K_{ij}^{ani} is the element in the i-th row and j-th column of K^{ani}, K^{cos} is the cosine similarity matrix, and N^{micr} represents the number of microbes.

Microbe–disease–drug knowledge graph

The microbe–disease–drug knowledge graph (Figure 1 (b)) comprises entity set E and a relationship set R. The entity set includes microbe, disease, and drug entities, while the relationship set consists of triplets $(h, r, t) \in E \times R \times E$, where h, r, and t denote the head entity, the relationship, and the tail entity, respectively. Each triplet represents a known fact; for example, (Ciprofloxacin, Association, Candida albicans) corresponds to the known fact that the microbe Candida albicans is inhibited by the drug ciprofloxacin.

Methods

We propose an prediction model, PCMDA (Figure 1), to predict microbe-drug associations. Extensive biological data has been integrated to form the integrated microbe-drug association dataset (Figure1 (a)). Representations of microbe and drug entities are learned through a neighborhood topology-aware knowledge graph module (Figure 1 (b)). We leverage a information-level attention mechanism to learn the preference features of microbes (Figure 1 (c)). Various biological connections are encoded and fused through a dual-gated network (Figure 1 (d)).

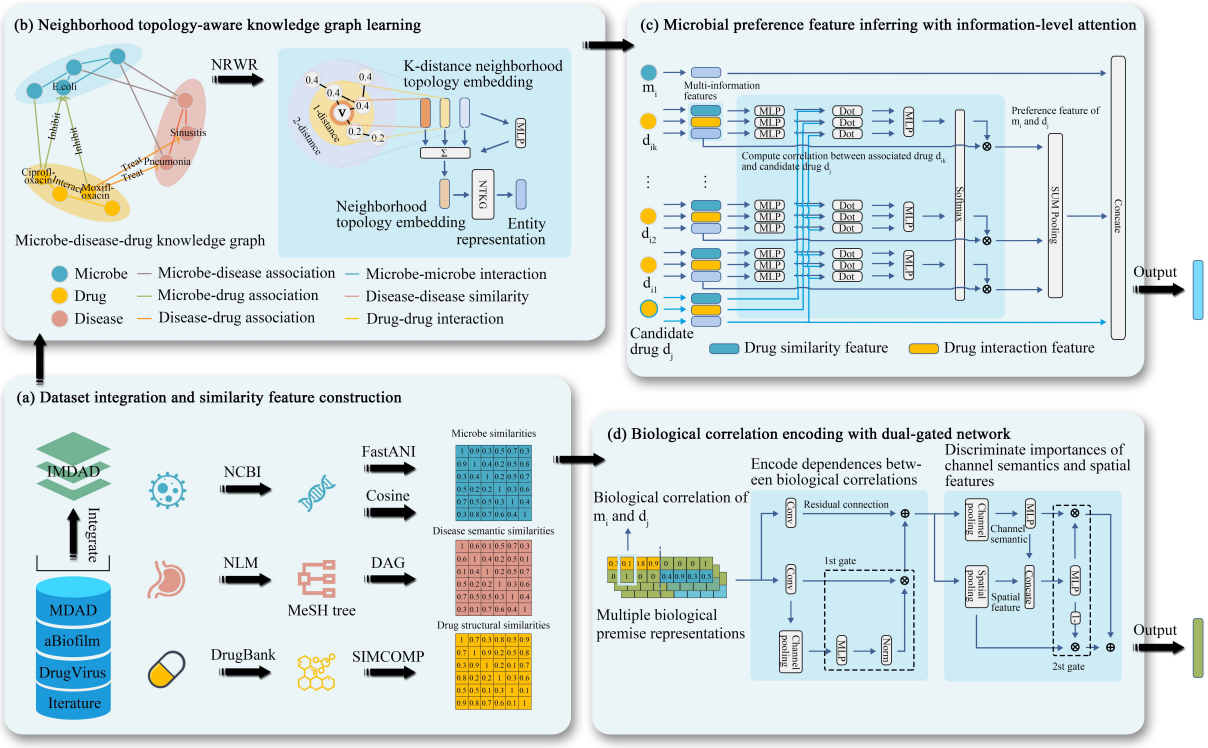


Figure 1: Overall architecture of PCMDA for predicting microbe–drug associations. a) Integrate microbe-drug association datasets and construct similarity features. b) Learn entity representations from multiple neighborhood topologies. c) Infer microbial preferences by information-level attention mechanism. d) Encode biological correlations based on a dual-gated network.

Neighborhood topology-aware knowledge graph learning based on multi-layer perceptron networks

Embedding encoding of entity neighborhood topology

The significance of neighboring entities and their topological connections[17] to a target entity is influenced by their shortest distances within the network. We have developed a strategy to encode the multi-scale topological neighborhood of entities based on the restart random walk algorithm within the k-distance neighborhood.

In this case, the k-distance neighborhood includes entities that are the shortest distance of k away from the target entity. The adjacency matrix of the microbe–disease–drug knowledge graph is denoted as

$$B^{mdd} = \begin{bmatrix} K^{drug} & B^{drug-micr} & B^{drug-dise} \\ (B^{drug-micr})^T & K^{micr} & B^{micr-dise} \\ (B^{drug-dise})^T & (B^{micr-dise})^T & K^{dise} \end{bmatrix}, \tag{2}$$

where K^{drug} and K^{dise} are drug (or disease) similarity matrices, $B^{drug-micr}$, $B^{drug-dise}$, and $B^{micr-dise}$ are drug–microbe (or drug–disease, microbe–disease) association matrices, and B^T denotes the transpose of B. Assuming a walker is at entity v at time t, then at time $t+1$, the state of the walker will be $\pi_v^{t+1} \in R^{1\times N}$,

$$\pi_v^{t+1} = \theta \cdot \hat{\pi}_v^t T + (1-\theta) \cdot \hat{\pi}_v^0, \tag{3}$$

where θ is the walk probability, $\hat{\pi_v^0}$ is a one-hot vector with the v-th element as 1, and T is the row-wise L_1 normalized adjacency matrix. To ensure the walker stays within the k-distance neighborhood, we mask and normalize π_v^{t+1} as follows:

$$\hat{\pi}_v^{t+1} = \frac{[sign((B^{mdd})_v^k) - sign((B^{mdd})_v^{k-1})] \cdot \pi_v^{t+1}}{[sign((B^{mdd})_v^k) - sign((B^{mdd})_v^{k-1})]^T \pi_v^{t+1}}, \tag{4}$$

where $(B^{mdd})_v^k$ represents the k-hop neighbors of an entity v, $sign(.)$ is the sign function, and $sign((B^{mdd})_v^k) - sign((B^{mdd})_v^{k-1})$ denotes the k-distance neighborhood of an entity.

After the walker's state converges, we estimate the topological proximity $\hat{\pi}_v$ between entity v and other entities in its k-distance neighborhood. The k-distance neighborhood topology embeddings of entity v are denoted as

$$X_v^{ent,k} = \hat{\pi}_v X^{ent}, \tag{5}$$

where X^{ent} is the entity embedding matrix initialized randomly. Since multi-scale topology embeddings contribute differently to entity representation, we calculate the importance vector α of entity embeddings under different distances,

$$\alpha = softmax((\|_{k=0}^{N^{nei}} X_v^{ent,k}) W_1^{nei} + b_1^{nei}), \tag{6}$$

where $\|$ denotes concatenation operation, N^{nei} is the number of neighborhoods, $X_v^{ent,0} = X_v^{ent}$, and W_1^{nei}, b_1^{nei} are the weight matrix and bias vector, respectively. As illustrated in Figure 1(b), the neighborhood topology embedding of entity v is formed as

$$\hat{X}_v^{ent} = \sum_{k=0}^{N^{nei}} \alpha_k \cdot X_v^{ent,k}. \tag{7}$$

Neighborhood topology-aware knowledge graph learning

On one hand, previous models of knowledge graphs[18–20] have traditionally focused on modeling known facts but often neglected the information related to the topology of entity neighborhoods. On the other hand, MLP-mixer[21] demonstrates strong capabilities in learning representations, particularly in exploring the relationship between topological embedding and relational embeddings. To address this gap, we propose neighborhood topology-aware knowledge graph learning based on MLP-mixer, as depicted in Figure 2.

Given an input triplet, to capture the connection between topological embeddings and

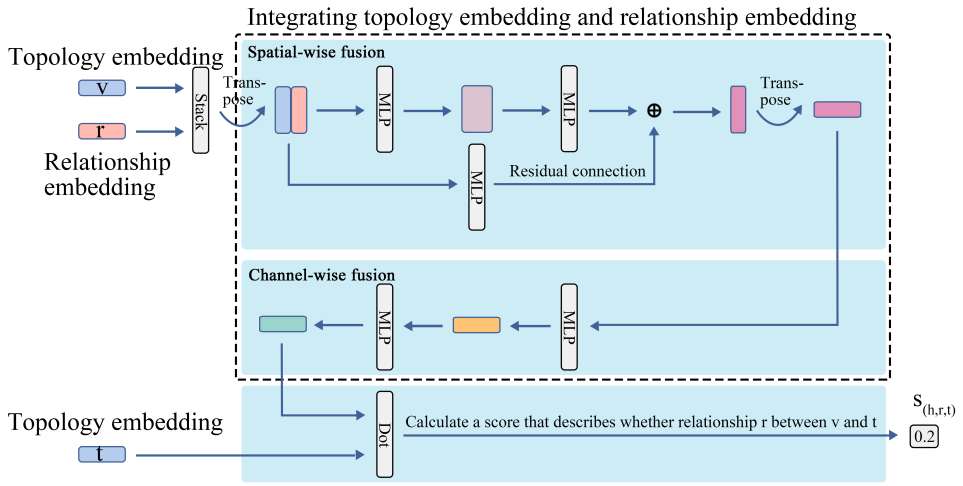


Figure 2: Process of neighborhood topology-aware knowledge graph learning.

relational embeddings in the spatial dimension, we concatenate the topological embedding of the head entity v with the embedding of the relation and perform spatial-wise fusion as follows

$$\hat{X}_{v,r}^{col} = W_2^{col} ReLU(W_1^{col} \begin{bmatrix} X_v^{ent} \\ X_r^{rel} \end{bmatrix} + b_1^{col}), \tag{8}$$

where X_r^{rel} is the r-th row in the relational embedding matrix, W_1^{col} and W_2^{col} are weight matrices, b_1^{col} and is a bias vector.

To mitigate the issue of gradient vanishing, a residual connection is introduced to create $X_{v,r}^{col}$,

$$X_{v,r}^{col} = \hat{X}_{v,r}^{col} + W^{res} \begin{bmatrix} X_v^{ent} \\ X_r^{rel} \end{bmatrix} + b^{res}, \tag{9}$$

where W^{res} and b^{res} are the weight matrix and bias vector, respectively. To further extract features of $X_{v,r}^{col}$ in the channel dimension, we apply channel-wise fusion to $X_{v,r}^{col}$ through fully connected layers, such that

$$X_{v,r}^{row} = ReLU(X_{v,r}^{col} W_1^{row} + b_1^{row}) W_2^{row} + b_2^{row}, \tag{10}$$

where W_1^{row}, W_2^{row} are weight matrices and b_1^{row}, b_2^{row} are bias vectors. The score of $s(v,r,t)$ is defined as follows,

$$s_{(v,r,t)} = sigmoid((X_{v,r}^{row})^T X_t^{ent}). \tag{11}$$

To ensure that entities with similar meanings are in close proximity to each other in the embedding space, we simultaneously optimize the representations of the entities and relations using a semantic smoothness loss[22] and cross-entropy loss,

$$\begin{aligned} L_k = -\frac{1}{\tau} \sum_{(v,r,t)\in\tau} (y_{(v,r,t)} \cdot log(s_{(v,r,t)}) + (1 - y_{(v,r,t)}) \cdot log(1- \\ s_{(v,r,t)})) + \frac{\lambda}{3|E|^2} \cdot \sum_{i}^{|E|} \sum_{j}^{|E|} (X_i^{ent} (X_j^{ent})^T - K_{ij}^{\phi(i)})^2, \end{aligned} \tag{12}$$

where τ is the set of training triplets, $y_{(v,r,t)}$ is the label of the triplet, λ is the weight of the semantic smoothness loss, and $\phi(i)$ denotes the type of the entity i.

Microbial preference feature inferring with information-level attention mechanism

In the microbe–disease–drug knowledge graph, the group of drugs associated with a microbe reflects its preferences in terms of drug associations, which are crucial for inferring the microbe's preference towards target drugs. Drawing inspiration from the deep interest network,[23] we have developed a microbial preference feature inferring (MPFI) module based on attention mechanisms to learn how microorganisms associate with candidate drugs, as depicted in Figure 1(c).

For a predicted pair of microbe entity m_i and candidate drug entity d_j, where their associated drugs are $d_{i_1} \cdots d_{i_k}$, we estimate the relevance between an associated drug d_{i_c} and the candidate drug d_j using a information-level attention mechanism. This integrates learned entity information, drug interaction information, and drug similarity information to

calculate attention scores β_{i_c},

$$\begin{aligned}\beta_{i_c} = [\hat{X}_{d_j}^{ent} W_1^{pre} (\hat{X}_{d_{i_c}}^{ent})^T, X_{d_j}^{int} W_2^{pre} (X_{d_{i_c}}^{int})^T, \\ K_{d_j}^{drug} W_3^{pre} (K_{d_{i_c}}^{drug})^T] W_4^{pre} + b_4^{pre},\end{aligned} \tag{13}$$

where $X_{d_j}^{int}$ is the corresponding row of d_j in drug interaction matrix, W_1^{pre}, W_2^{pre}, and W_3^{pre} are weight matrices, W_4^{pre} is a weight vector, and b_4^{pre} is a bias vector. Considering that d_j may be a new drug with fewer associations, we normalize the attention scores using softmax,

$$\hat{\beta} = softmax([1, \beta_{i_1}, \cdots, \beta_{i_k}]), \tag{14}$$

where $\hat{\beta}_{i_c}$ reflects the degree of relevance between d_j and d_{i_c}. By aggregating the features of associated drugs related to the candidate drug, we derive the preference feature $X_{m_i}^{pre}$ of m_i toward d_j,

$$X_{m_i}^{pre} = \sum_{c=1}^{k} \hat{\beta}_{i_c} X_{d_{i_c}}^{ent}. \tag{15}$$

This preference feature encapsulates the microbe's inclination to associate with the candidate drug. Finally, by concatenating the features of the microbe m_i, the candidate drug d_j, and the preference feature, we obtain the final output $\hat{X}_{(m_i,d_j)}^{pre}$,

$$\hat{X}_{(m_i,d_j)}^{pre} = [\hat{X}_{m_i}^{ent}, X_{m_i}^{pre}, \hat{X}_{d_j}^{ent}]. \tag{16}$$

Biological correlation encoding with dual-gated network

Biological premises of microbe–drug associations

Biological premises provide insights into the associations between microbes and drugs from a biological perspective. In Table 1, we outline the common premises underlying microbe–drug associations.

Table 1: Premises of common microbe–drug associations.

Premise	Description	Type
1	Drugs (microbes) with similar structures are associated with the same microbes (drugs)	Similarity-based
2	Drugs (microbes) with similar associated microbes (drugs) are associated with the same microbes (drugs)	Similarity-based
3	Interacting drugs (microbes) are associated with the same microbes (drugs)	Interaction-based

We construct the feature corresponding to premise 1 as

$$X^{pme_1}_{(m_i,d_j)} = \begin{bmatrix} K^{drug}_{d_j} & B^{drug-micr}_{d_j} \\ (B^{drug-micr})^T_{m_i} & K^{micr}_{m_i} \end{bmatrix}, \tag{17}$$

where each column c in $X^{pme_1}_{(m_d,d_j)}$ indicates whether there exists a biological correlation between d_j and m_i (i.e., d_j is similar to d_c and d_c is associated with m_i), with $1 \leq c \leq N^{drug}$. The drug similarity matrix K^{asso_1} is calculated based on the drug–microbe association matrix using the cosine similarity function,

$$K^{asso_1} = norm_2(B^{drug-micr})norm_2(B^{drug-micr})^T, \tag{18}$$

where $norm_2$ denotes row-wise L_2 normalization. Similarly, the features corresponding to premise 2 and 3 are represented as $X^{pme_2}_{(m_i,d_j)}$ and $X^{pme_3}_{(m_i,d_j)}$, respectively. These features are stacked by channel to integrate the aforementioned biological premises, resulting in the final premise feature $X^{pme}_{(m_i,d_j)} \in R^{3\times 2\times (N^{drug}+N^{micr})}$.

Biological correlation encoding

To encode the hidden biological correlations in $X^{pme}_{m_i,d_j}$, we propose a dual-gated network for biological correlation encoding (BCE), illustrated in Figure 1(d). First, we fuse features

across different premises using convolution,

$$X^{fus}_{(m_i,d_j)} = ReLU(X^{pme}_{(m_i,d_j)} * \omega), \tag{19}$$

where $*$ denotes the convolution operation and ω represents N^{flt} filters of size $3\times2\times1$. Next, a gated network is employed to capture spatial interactions among features,[24] incorporating dependencies of biological correlations across different positions. The gating signal is learned by fusing features extracted through pooling operations using a fully connected network, with detailed information supplemented in a residual manner,

$$\begin{aligned}\hat{X}^{fus}_{(m_i,d_j)} = norm_2(sum_1(X^{fus}_{(m_i,d_j)})W^{gat_1} + b^{gat_1}) \\ \odot X^{fus}_{(m_i,d_j)} + X^{fus}_{(m_i,d_j)},\end{aligned} \tag{20}$$

where sum_1 denotes a channel-level sum pooling operation, $\odot$ represents the Hadamard product operation, W^{gat_1} is a weight matrix, and b^{gat_1} is a bias vector. Channel pooling integrates semantics across different channels, while spatial pooling fuses features across different positions. We then perform channel and spatial pooling on $\hat{X}^{fus}_{(m_i,d_j)}$ to obtain $X^{cha}_{(m_i,d_j)}$ and $X^{spa}_{(m_i),d_j}$, respectively,

$$X^{spa}_{(m_i,d_j)} = sum_1(\hat{X}^{fus}_{(m_i,d_j)})W^{spa} + b^{spa}, \tag{21}$$

$$X^{cha}_{(m_i,d_j)} = sum_2(\hat{X}^{fus}_{(m_i,d_j)}), \tag{22}$$

where sum_2 denotes a spatial-level sum pooling, W^{spa} is a weight matrix, and b^{spa} is a bias vector. Finally, a gated network distinguishes the importance of channel semantics and spatial features and obtains biological correlation feature,

$$g_{(m_i,d_j)} = sigmoid([X^{cha}_{(m_i,d_j)}, X^{spa}_{(m_i,d_j)}]W^{gat_2} + b^{gat_2}), \tag{23}$$

$$\hat{X}^{pme}_{(m_i,d_j)} = g_{(m_i,d_j)} \odot X^{cha}_{(m_i,d_j)} + (1 - g_{(m_i,d_j)}) \odot X^{spa}_{(m_i,d_j)}, \tag{24}$$

where W^{gat_2} is a weight matrix and b^{gat_2} is a bias vector.

Representation integration and optimization

Microbial preference feature reflects the correlation tendency of m_i toward d_j, while biological correlation feature represents the dependencies of different biological correlations under various biological premises. After fusing these features using a fully connected network, the probability distribution $\hat{p}_{(m_i,d_j)}$ of association between m_i and d_j is calculated as follows,

$$p_{(m_i,d_j)} = ReLU([\hat{X}^{pre}_{(m_i,d_j)} \hat{X}^{pme}_{(m_i,d_j)}]W_1^{pred} + b_1^{pred}), \tag{25}$$

$$\hat{p}_{(m_i,d_j)} = softmax(p_{(m_i,d_j)}W_2^{pred} + b_2^{pred}), \tag{26}$$

where W_1^{pred}, W_2^{pred} are weight matrices, and b_1^{pred}, b_2^{pred} are bias vectors. The loss function for association prediction is defined as

$$\begin{aligned} L_p = -\frac{1}{|\wedge|} \sum_{(m_i,d_j)\in\wedge} & (y_{(m_i,d_j)} \cdot log(\hat{p}_{(m_i,d_j)}) + \\ & (1 - y_{(m_i,d_j)}) \cdot log(1 - \hat{p}_{(m_i,d_j)})), \end{aligned} \tag{27}$$

where $\wedge$ is the training set and $y_{(m_i,d_j)}$ is the true label.

Experiments and Discussions

Evaluation metrics and parameter settings

All microbe–drug associations observed and unobserved in IMDAD serve as positive and negative examples, respectively. To maintain model generalization, the dataset is randomly partitioned into 80% for training, 10% for validation, and 10% for testing. Evaluation metrics

such as AUC, AUPR, and F1[25,26] are typically used to assess model performance by ranking prediction scores. The average AUC and AUPR metrics, as discussed in,[27] are employed to evaluate how well models predict candidate microbes related to drugs. For simplicity, we will continue to refer to these metrics as AUC and AUPR. Given that biologists often prioritize highly ranked candidate microbes for wet experiments, we calculate the average recall rate across all drugs under various top h candidate microbe settings. PCMDA was performed on an RTX 2070 server using PyTorch 1.12.1,[28] with model optimization carried out using the Adam algorithm.[29] To avoid overfitting, we used early stopping techniques. The hyperparameters for the model were set as follows: the feature dimension was fixed at 128, and the learning rate and weight decay were set to 1e-3 and 1e-4, respectively. To improve entity representations, we set the walk probability in the neighborhood-restarted random walk to 0.8, and the number of neighborhoods to 2. The weight assigned to the semantic smoothing loss, which affects entity distributions, was set to 1. In the biological correlation encoding component, we utilized 256 filters to ensure effective extraction of relevant knowledge.

Comparison with other methods

We conducted experiments on the IMDAD dataset and compared PCMDA with several state-of-the-art computational methods using the same data separation. The comparison methods are briefly described as follows:

- ConvE:[19] Predicts scores of triplets in knowledge graphs using convolutional neural networks.
- GIN:[30] Graph isomorphism network learns discriminative entity representations based on the Weisfeiler–Lehman graph isomorphism test.
- GSAMDA:[31] Predicts microbe–drug associations using sparse autoencoders and graph attention networks.

- GACNNMDA:[32] Utilizes graph attention networks and convolutional networks for microbe–drug association prediction.
- MKGCN:[33] Detects microbe–drug associations based on multiple kernel fusion and GCN.
- GNAEMDA:[5] Learns association distributions of microbes and drugs using a graph autoencoder with feature normalization.
- NGMDA:[6] Infers microbe–drug associations based on graph attention networks and graph transformer.

Table 2: AUCs, AUPRs, and F1 values of different methods (the best results are marked in bold)

Network	**AUC**	**AUPR**	**F1**
ConvE	91.84%	74.57%	82.31%
GIN	84.76%	62.29%	71.81%
GSAMDA	94.33%	75.26%	83.72%
GACNNMDA	92.30%	75.87%	83.28%
MKGCN	91.98%	80.76%	86.00%
GNAEMDA	93.41%	77.68%	84.82%
NGMDA	94.56%	80.53%	86.98%
PCMDA	**96.19%**	**83.01%**	**89.12%**

We conducted the experiments five times using different seeds, and Table 2 presents the averaged AUCs, AUPRs, and F1 values of various prediction approaches on the IMDAD dataset. According to the experimental results, PCMDA achieves the highest F1 score of 89.12%, which is 2.13% higher than the second-best performing method, NGMDA. GSAMDA and GACNNMDA demonstrate better performance compared to baseline approaches (ConvE and GIN), although they exhibit relatively poorer performance compared to other methods possibly due to their simpler model structures. MKGCN and GNAEMDA, which do not fully consider the diverse connections between microbes and drugs, achieve slightly lower performance than NGMDA. While NGMDA effectively models entity heterogeneity and relationships, capturing connections between different entities, it does not extensively explore

rich biological hypotheses or integrate multisource information such as disease details and gene sequence consistency.

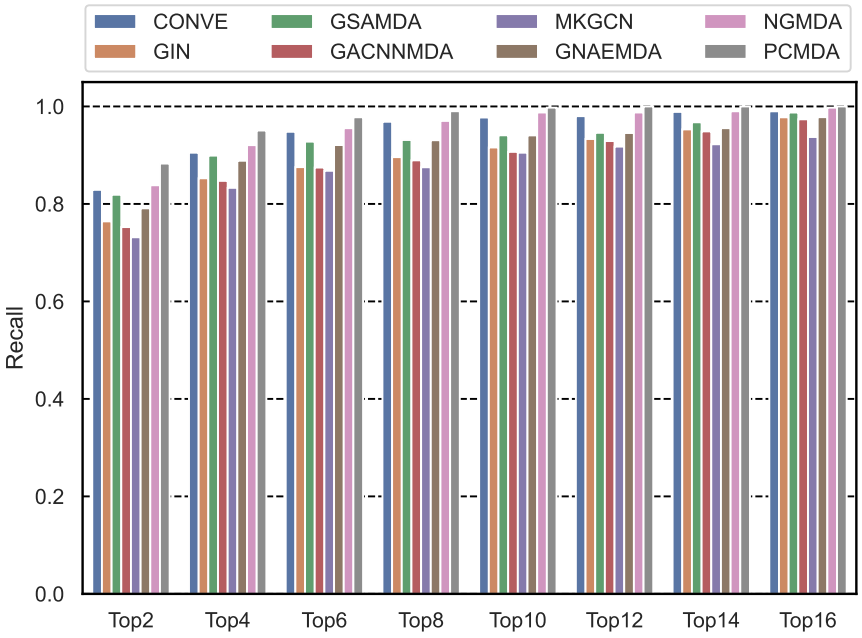


Figure 3: Average recalls of drugs at different top h cutoffs

Figure 3 illustrates the average recall rates of various drugs across different cutoffs, highlighting PCMDA as the top performer. As the cutoff value (h) increases from 2 to 6, PCMDA achieves recall rates of 88.28%, 95.06%, and 97.78%, respectively. The second best is NGMDA with recall rates of 83.83%, 92.06%, and 95.54%, respectively. ConvE outperforms GSAMDA with recall rates of 82.89%, 90.51%, and 94.80%, respectively, compared to 81.88%, 89.91%, and 92.78% for GSAMDA. GACNNMDA (75.25%, 84.73%, and 87.46%) and GIN (76.42%, 85.26%, and 87.55%) perform relatively poorly compared to GANEMDA, which achieves recall rates of 79.10%, 88.86%, and 92.08%, respectively. MKGCN achieves the lowest recalls with rates of 73.16%, 83.30%, and 86.80%, respectively.

Ablation experiments

We conducted an ablation experiment (Table 3) to assess the effectiveness of several components: neighborhood topology encoding strategy (NTES), topology-aware knowledge graph

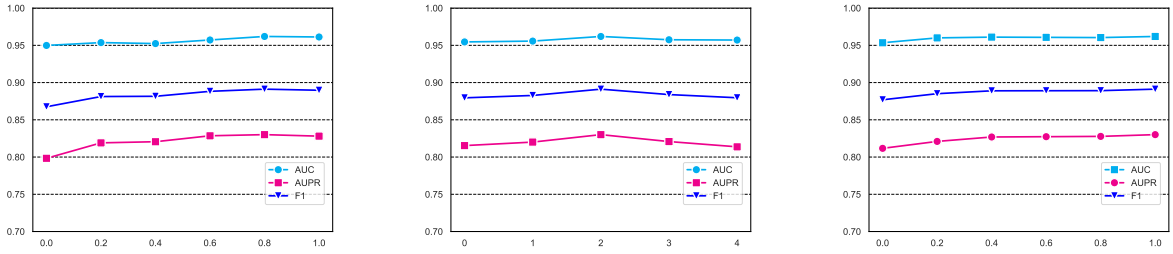
learning (TAKG), BCE, and MPFI. Among these components, TAKG showed the most significant impact on the model's performance, leading to an 8.14% decrease in the F1 score. This outcome suggests that TAKG effectively learns entities, relations, and observed facts within the knowledge graph. For PCMDA without MPFI, there was a decrease of 3.42% in the F1 score, indicating that inferring microbial preferences regarding candidate drugs can enhance prediction performance. Removing NTES from PCMDA resulted in a 1.00% decrease in the F1 score, highlighting the importance of neighborhood topology information for accurate entity representation. Additionally, BCE contributes biologically relevant assumptions about microbe–drug associations, resulting in a 0.66% increase in the F1 score.

Table 3: Results of ablation experiments

Network	**AUC**	**AUPR**	**F1**
PCMDA	96.19%	83.01%	89.12%
PCMDA w/o NTES	95.86%	81.54%	88.12%
PCMDA w/o TAKG	93.10%	71.65%	80.98%
PCMDA w/o MPFI	95.37%	77.82%	85.70%
PCMDA w/o BCE	96.03%	82.00%	88.46%

Parameter sensitivity

To evaluate the impact of hyperparameters on model performance, we conducted a sensitivity analysis on three key parameters: walk probability θ, number of neighbors N^{nei}, and semantic smoothing loss weight λ, as depicted in Figure 4. The walk probability determines the direction of the walker in the knowledge graph, and we tested values from $\{0, 0.2, \ldots, 1\}$. PCMDA achieved its highest F1 score when θ was set to 0.8 (Figure 4(a)). This suggests that directing the walker more towards existing connections in the graph optimizes model performance. The parameter N^{nei}, representing the number of neighboring entities considered, was fine-tuned from 0 to 4 with an interval of 1. Figure 4(b) shows that the optimal performance was observed when N^{nei} was set to 2, indicating that entities in closer proximity to the target entity have a more significant influence on predictions. Semantic smoothing



(a) Parameter study of θ (b) Parameter study of N^{nei} (c) Parameter study of λ

Figure 4: Parameter sensitivity study on IMDAD

loss weight λ regulates the model's ability to capture semantic relationships between entities and was varied over $\{0, 0.2, \ldots, 1\}$. According to Figure 4(c), setting λ to 1 yielded the best results, emphasizing the importance of enhancing entity representations in the embedding space to improve predictive accuracy.

Case studies on three drugs

To evaluate the model's ability to identify microbes related to drugs, we conducted case studies involving ciprofloxacin, moxifloxacin, and vancomycin. Ciprofloxacin and moxifloxacin are fluoroquinolone antibiotics used to treat bacterial infections such as skin infections and pneumonia.[34,35] Vancomycin, a glycopeptide antibiotic, is used to treat bacterial infections including skin, joint, and bloodstream infections.[36] The experimental process proceeded as follows: We trained the model using all known and previously unseen associations between microbes and drugs. Then, we ranked the candidate microbes related to drugs based on their scores in descending order. To assess the model's effectiveness in identifying candidate microbes related to drugs, we verified the top 20 candidates using IMDAD and referenced literature sources.

Regarding the top 20 candidate microbes related to ciprofloxacin in Table 3, 15 ones are documented in IMDAD and 5 ones are supported by literature. This demonstrates the method's ability to infer associations between microbes and drugs. Specifically, Strepto-

coccus pneumoniae, Salmonella enterica, and Clostridium perfringens are reported to be sensitive to ciprofloxacin,[37–39] while Acinetobacter baumannii and Enterococcus faecalis are known for their resistance to ciprofloxacin.[40,41] For the candidate microbes related to moxifloxacin in Table 4, nine candidates are identified in IMDAD and seven are validated by literature sources. Regarding the candidate microbes for vancomycin in Table 5, IMDAD lists 13 candidates, with three confirmed by literature sources. Among a total of 60 candidates across these studies, eight remain unconfirmed, indicating a lack of supporting evidence for those associations. Overall, these experimental findings illustrate that our model effectively predicts candidate microbes related to drugs.

Table 4: The top-20 candidate microbes of Ciprofloxacin.

Rank	Microbe name	Evidence	Rank	Microbe name	Evidence
1	Escherichia coli	IMDAD	11	Proteus mirabilis	IMDAD
2	Haemophilus influenzae	IMDAD	12	Stenotrophomonas maltophilia	IMDAD
3	Staphylococcus aureus	IMDAD	13	Streptococcus pneumoniae	PMID:26100702
4	Pseudomonas aeruginosa	IMDAD	14	Listeria monocytogenes	IMDAD
5	Providencia stuartii	IMDAD	15	Streptococcus mutans	IMDAD
6	Morganella morganii	IMDAD	16	Staphylococcus epidermidis	IMDAD
7	Mycobacterium tuberculosis	IMDAD	17	Salmonella enterica	PMID:26933017
8	Klebsiella pneumoniae	IMDAD	18	Acinetobacter baumannii	PMID:25705272
9	Proteus vulgaris	IMDAD	19	Enterococcus faecalis	PMID:27790716
10	Bacillus subtilis	IMDAD	20	Clostridium perfringens	PMID:29978055

Prediction of novel microbe–drug associations

To aid biologists in conducting drug-related experiments, our model predicts candidate microbes for all drugs. The top 20 microbe candidates for each drug are listed in Supplementary File SF2.

Table 5: The top-20 candidate microbes of Moxifloxacin.

Rank	Microbe name	Evidence	Rank	Microbe name	Evidence
1	Pseudomonas aeruginosa	IMDAD	11	Bacillus subtilis	PMID:30036828
2	Stenotrophomonas maltophilia	IMDAD	12	Staphylococcus epidermidis	PMID:11249827
3	Staphylococcus aureus	IMDAD	13	Enterococcus faecalis	PMID:31763048
4	Haemophilus influenzae	IMDAD	14	Proteus mirabilis	Unconfirmed
5	Mycobacterium avium	IMDAD	15	Candida glabrata	Unconfirmed
6	Escherichia coli	IMDAD	16	Micrococcus luteus	Unconfirmed
7	Listeria monocytogenes	IMDAD	17	Human immunodeficiency virus	PMID:18441333
8	Streptococcus pneumoniae	IMDAD	18	Mycobacterium tuberculosis	PMID:35975988
9	Klebsiella pneumoniae	IMDAD	19	Burkholderia cepacia	Unconfirmed
10	Streptococcus mutans	PMID:29160117	20	Clostridium perfringens	PMID:29486533

Table 6: The top-20 candidate microbes of Vancomycin.

Rank	Microbe name	Evidence	Rank	Microbe name	Evidence
1	Staphylococcus aureus	IMDAD	11	Listeria monocytogenes	IMDAD
2	Staphylococcus epidermidis	IMDAD	12	Streptococcus mutans	IMDAD
3	Enterococcus faecalis	IMDAD	13	Pseudomonas aeruginosa	IMDAD
4	Staphylococcus capitis	IMDAD	14	Streptococcus pneumoniae	PMID:10376600
5	Candida tropicalis	IMDAD	15	Escherichia coli	PMID:33468474
6	Enterococcus faecium	IMDAD	16	Haemophilus influenzae	Unconfirmed
7	Staphylococcus caprae	IMDAD	17	Bacillus subtilis	PMID:14165485
8	Staphylococcus chromogenes	IMDAD	18	Stenotrophomonas maltophilia	Unconfirmed
9	Amycolatopsis orientalis	IMDAD	19	Salmonella enterica	Unconfirmed
10	Staphylococcus cohnii	IMDAD	20	Cryptococcus neoformans	Unconfirmed

Conclusions

We proposed a method to integrate the neighborhood topologies of each entity covering multiple scopes, the diverse relationship semantics, and the pairwise entity features for inferring the candidate microbes for the interested drugs. Random walks on the constructed microbe-disease-drug knowledge graph facilitated formulation of the entity topologies with multiple scales. The neighborhood topologies and the relationship semantics of the microbes, diseases, and drugs were deeply integrated from both the channel and spatial perspectives. The designed information-level attention was helpful for evaluating the relevance between the group of drugs that prefer to associate with a microbe and a single candidate drug. The constucted dual-gated network was able to adaptively fuse the pairwise entity features from multiple biological perspectives. The comparison experiments showed that PCMDA achieved higher AUC and AUPR than the state-of-the-art prediction methods. The ablation experiment results confirmed the effectiveness of neighborhood topology-aware knowledge graph learning and microbial preference feature inferring, and the parameter sensitivity analysis demonstrated the impact of hyper-parameter selection on prediction performance. The case studies on three drugs highlighted PCMDA's ability in discovering the reliable candidate microbes for the drugs.

Data and Software Availability

The source code and datasets are freely available at https://github.com/pingxuan-hlju/PCMDA.

Acknowledgement

This work was supported by Natural Science Foundation of China (62372282, 62172143); Natural Science Foundation of Guangdong Province (2024A1515010176); Natural Science

Foundation of Heilongjiang Province (LH2023F044); STU Scientific Research Initiation Grant (NTF22032).

Supporting Information Available

- The 452 microbe-drug associations extracted from published literature (SF1.xlsx)
- The top 20 potential candidates for 1209 drugs predicted by our method (ST2.xlsx)

References

(1) Bowe, B.; Xie, Y.; Al-Aly, Z. Postacute sequelae of COVID-19 at 2 years. *Nature Medicine* **2023**, *29*, 2347–2357.

(2) Mulder, D.; Aarts, E.; Arias Vasquez, A.; others A systematic review exploring the association between the human gut microbiota and brain connectivity in health and disease. *Mol Psychiatry* **2023**, *9*, 5037–5061.

(3) V. Kumbhare, S.; Pedroso, I.; A. Ugalde, J.; others Drug and gut microbe relationships: Moving beyond antibiotics. *Drug Discovery Today* **2023**, *28*, 103797.

(4) Long, Y.; Wu, M.; Liu, Y.; others Ensembling graph attention networks for human microbe-drug association prediction. *Bioinformatics* **2020**, *36*, i779–i786.

(5) Huang, H.; Sun, Y.; Lan, M.; others GNAEMDA: Microbe-Drug Associations Prediction on Graph Normalized Convolutional Network. *IEEE Journal of Biomedical and Health Informatics* **2023**, *27*, 1635–1643.

(6) Xuan, P.; Gu, J.; Cui, H.; others Multi-scale topology and position feature learning and relationship-aware graph reasoning for prediction of drug-related microbes. *Bioinformatics* **2024**, *40*, btae025.

(7) Sun, Y.-Z.; Zhang, D.-H.; Cai, S.-B.; others MDAD: a special resource for microbe-drug associations. *Frontiers in cellular and infection microbiology* **2018**, *8*, 424.

(8) Rajput, A.; Thakur, A.; Sharma, S.; others aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic acids research* **2018**, *46*, D894–D900.

(9) I. Andersen, P.; Ianevski, A.; Lysvand, H.; others Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseases* **2020**, *93*, 268–276.

(10) Qi, C.; Cai, Y.; Qian, K.; others gutMDisorder v2.0: a comprehensive database for dysbiosis of gut microbiota in phenotypes and interventions. *Nucleic Acids Res* **2023**, *51*, D717–D722.

(11) Ma, W.; Zhang, L.; Zeng, P.; others An analysis of human microbe–disease associations. *Briefings in Bioinformatics* **2016**, *18*, 85–97.

(12) Skoufos, G.; Kardaras, F. S.; Alexiou, A.; others Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Research* **2020**, *49*, D1328–D1333.

(13) Davis, A. P.; Grondin, C. J.; Johnson, R. J.; others The comparative toxicogenomics database: update 2019. *Nucleic Acids Research* **2018**, *47*, D948–D954.

(14) Knox, C.; Wilson, M.; Klinger, C. M.; others DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research* **2024**, *52*, D1265–D1275.

(15) Hattori, M.; Tanaka, N.; Kanehisa, M.; others SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic acids research* **2010**, *38*, W652–W656.

(16) Jain, C.; M. Rodriguez-R, L.; M.Phillippy, A.; others High throughput ANI analysis

of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **2018**, *9*.

(17) Cai, X.; Xia, L.; Ren, X.; others How Expressive are Graph Neural Networks in Recommendation? Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York, NY, USA, 2023.

(18) Bordes, A.; Usunier, N.; Garcia-Durán, A.; others Translating embeddings for modeling multi-relational data. Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2013.

(19) Dettmers, T.; Minervini, P.; Stenetorp, P.; others Convolutional 2D knowledge graph embeddings. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, Louisiana, USA, 2018.

(20) Zhang, Z.; Wang, J.; Ye, J.; others Rethinking Graph Convolutional Networks in Knowledge Graph Completion. Proceedings of the ACM Web Conference 2022. New York, NY, USA, 2022.

(21) Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; others MLP-mixer: an all-MLP architecture for vision. Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2024.

(22) Guo, S.; Wang, Q.; Wang, B.; others Semantically Smooth Knowledge Graph Embedding. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015.

(23) Zhou, G.; Mou, N.; Fan, Y.; others Deep interest evolution network for click-through rate prediction. Proceedings of the Thirty-Third AAAI Conference on Artificial Intelli-

gence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu, Hawaii, USA, 2019.

(24) Rao, Y.; Zhao, W.; Tang, Y.; others HorNet: efficient high-order spatial interactions with recursive gated convolutions. Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2024.

(25) Huang, J.; Ling, C. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* **2005**, *17*, 299–310.

(26) Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **2015**, *10*.

(27) Xuan, P.; Cao, Y.; Zhang, T.; others Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* **2019**, *35*, 4108–4119.

(28) Paszke, A.; Gross, S.; Massa, F.; others PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2019.

(29) P. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **2014**,

(30) Xu, K.; Hu, W.; Leskovec, J.; others How Powerful are Graph Neural Networks? International Conference on Learning Representations. 2019.

(31) Tan, Y.; Zou, J.; Kuang, L.; others GSAMDA: a computational model for predicting potential microbe-drug associations based on graph attention network and sparse autoencoder. *BMC bioinformatics* **2022**, *23*, 492.

(32) Ma, Q.; Tan, Y.; Wang, L. GACNNMDA: a computational model for predicting potential human microbe-drug associations based on graph attention network and CNN-based classifier. *BMC bioinformatics* **2023**, *24*, 35.

(33) Yang, H.; Ding, Y.; Tang, J.; others Inferring human microbe–drug associations via multiple kernel fusion on graph neural network. *Knowledge-Based Systems* **2022**, *238*, 107888.

(34) Thai, T.; H. Salisbury, B.; M. Zito, P. Ciprofloxacin. *In StatPearls* **2023**,

(35) M. Keating, G.; J. Scott, L. Moxifloxacin: a review of its use in the management of bacterial infections. *Drugs* **2004**, *64*, 2347–2377.

(36) Bruniera, F. R.; Ferreira, F. M.; Saviolli, L. R. M.; others The use of vancomycin with its therapeutic and adverse effects: a review. *European Review for Medical and Pharmacological Sciences* **2015**, *19*, 694–700.

(37) Dridi, B.; Lupien, A.; G. Bergeron, M.; others Differences in antibiotic-induced oxidative stress responses between laboratory and clinical isolates of Streptococcus pneumoniae. *Antimicrobial agents and chemotherapy* **2015**, *59*.

(38) Eibach, D.; M Al-Emran, H.; Myriam Dekker, D.; others The Emergence of Reduced Ciprofloxacin Susceptibility in Salmonella enterica Causing Bloodstream Infections in Rural Ghana. *Clinical Infectious Diseases* **2016**, *62*, S32–S36.

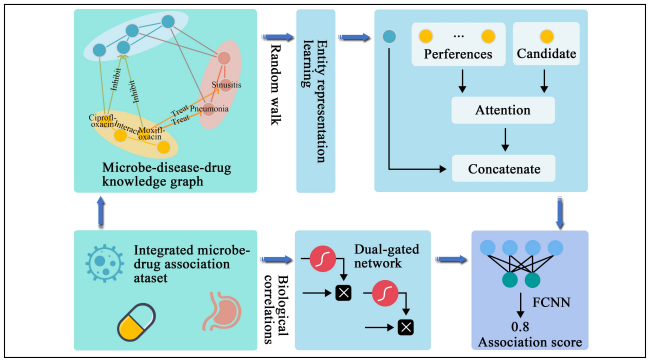
(39) Hamza, D.; Dorgham, S.; Hakim, A. Toxinotyping and Antimicrobial Resistance of Clostridium Perfringens Isolated from Processed Chicken Meat Products. *J Vet Res* **2017**, *61*, 53–58.

(40) Maleki, M.-H.; Azizi Jalilian, F.; Khayat, H.; others Detection of highly ciprofloxacin resistance acinetobacter baumannii isolated from patients with burn wound infections in presence and absence of efflux pump inhibitor. *Maedica (Bucur)* **2014**, *9*, 2.

(41) Kim, M.-C.; Woo, G.-J. Characterization of antimicrobial resistance and quinolone resistance factors in high-level ciprofloxacin-resistant Enterococcus faecalis and Ente-

rococcus faecium isolates obtained from fresh produce and fecal samples of patients. *J Sci Food Agric* **2017**, *97*, 2858–2864.

TOC Graphic


Inhibit
Inhibit
Sinusitis
Treat
Treat
Pneumonia
Ciprofl-oxacin
Isoniaz-id
Moxifl-oxacin
Microbe-disease-drug knowledge graph
Random walk
Entity representation learning
...
Perferences
Candidate
Attention
Concatenate
Integrated microbe-drug association ataset
Biological correlations
Dual-gated network
FCNN
0.8
Association score