Briefings in Bioinformatics, 2023, **24(2)**, 1–19
https://doi.org/10.1093/bib/bbac634
Advance access publication date 30 January 2023
Problem Solving Protocol

Predicting microbe–drug associations with structure-enhanced contrastive learning and self-paced negative sampling strategy

Zhen Tian, Yue Yu, Haichuan Fang, Weixin Xie and Maozu Guo

Corresponding author. Maozu Guo School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, 100044 Beijing, China. Tel:+86-010-61209266; E-mail: guomaozu@bucea.edu.cn

Abstract

Motivation: Predicting the associations between human microbes and drugs (MDAs) is one critical step in drug development and precision medicine areas. Since discovering these associations through wet experiments is time-consuming and labor-intensive, computational methods have already been an effective way to tackle this problem. Recently, graph contrastive learning (GCL) approaches have shown great advantages in learning the embeddings of nodes from heterogeneous biological graphs (HBGs). However, most GCL-based approaches don't fully capture the rich structure information in HBGs. Besides, fewer MDA prediction methods could screen out the most informative negative samples for effectively training the classifier. Therefore, it still needs to improve the accuracy of MDA predictions.
Results: In this study, we propose a novel approach that employs the Structure-enhanced Contrastive learning and Self-paced negative sampling strategy for Microbe-Drug Association predictions (SCSMDA). Firstly, SCSMDA constructs the similarity networks of microbes and drugs, as well as their different meta-path-induced networks. Then SCSMDA employs the representations of microbes and drugs learned from meta-path-induced networks to enhance their embeddings learned from the similarity networks by the contrastive learning strategy. After that, we adopt the self-paced negative sampling strategy to select the most informative negative samples to train the MLP classifier. Lastly, SCSMDA predicts the potential microbe–drug associations with the trained MLP classifier. The embeddings of microbes and drugs learning from the similarity networks are enhanced with the contrastive learning strategy, which could obtain their discriminative representations. Extensive results on three public datasets indicate that SCSMDA significantly outperforms other baseline methods on the MDA prediction task. Case studies for two common drugs could further demonstrate the effectiveness of SCSMDA in finding novel MDA associations.
Availability: The source code is publicly available on GitHub https://github.com/Yue-Yuu/SCSMDA-master.

Keywords: structure-enhanced contrastive learning, self-paced negative sampling, microbe–drug association prediction.

Introduction

Microbe or microorganism is a category of microscopic living organisms that have close interactions with human hosts. Generally, one microbe community mainly contains bacteria, viruses, protozoa and fungi [1]. Recent studies have shown that microbe communities usually play significant roles in human health, such as facilitating metabolism [2], producing essential vitamins [3] and protecting against invasion from pathogens [4]. However, the imbalance or dysbiosis of microbe communities may also cause some common infectious diseases such as obesity [5], diabetes [6] and even cancer [7]. Therefore, discovering the relationship between microbes and drugs is one essential problem for precision medicine [8–10].

Since inferring these associations with conventional wet-lab experiments is time-consuming, computational methods have already been proposed to tackle this problem. Moreover, with the increasing availability of various data sources related to microbes and drugs, these computational-based approaches have gained remarkable success [11]. For example, Zhu [12] raised HMDAKATZ method that predicted the potential associations based on the microbe–drug heterogeneous network. Long proposed GCNMDA model that first measured the similarity between microbes and drugs and then employed the conditional random field-based framework to learn their deep representations [13]. HNERMDA [14] constructed the microbe–drug heterogeneous network and adopted the metapath2vec model to learn the low-dimensional

Zhen Tian, PhD (Harbin Institute of Technology), is a lecturer at the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China. His current research interests include computational biology, complex network analysis and data mining.
Yue Yu is currently studying toward the Master Degree of Computer Science and Technology in Zhengzhou University, Zhengzhou, China. His research interests include knowledge graph embedding, bioinformatics and deep learning.
Haichuan Fang is currently working toward the Master Degree of Engineering in Zhengzhou University, Zhengzhou, China. His research interests include knowledge graph embedding, bioinformatics and deep learning.
Weixin Xie. Weixin Xie, Ph.D. (Harbin Engineering University, Harbin, China). Her research focuses on biomedical informatics, deep learning and text mining.
Maozu Guo is a professor at the College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China. He received the PhD degree in Computer Science and Technology from Harbin Institute of Technology. His research interests include bioinformatics, machine learning and data mining.
Received: November 8, 2022. **Revised:** December 19, 2022. **Accepted:** December 29, 2022

embeddings. EGATMDA [15] aimed to fully utilize the multisource of microbes and drugs to discover their association relationships. This model could learn the importance of different heterogeneous networks with graph-level attention mechanism and then obtain a deep representation of microbes and drugs. Meanwhile, Graph2MDA [16] employed the variational graph autoencoder to obtain the informative and interpretable latent representations for microbes and drugs based on their multimodal attributed graphs. Besides, MKGCN [17] first extracted the features of microbes and drugs at different graph convolutional network (GCN) layers and then predicted the microbe–drug association with multiple kernel matrices. However, these approaches above may have some weaknesses. For example, HMDAKATZ only adopted simple metrics to evaluate the association strengths between microbes and drugs, while GCNMDA and EGATMDA only selected negative samples in a random manner, which ignored the effects of different negative samples on the prediction model. Meanwhile, MKGCN couldn't fully capture the complex structure and rich semantics between nodes in the heterogeneous networks.

Recently, self-supervised learning approaches have attracted considerable attention because they provided novel insights into decreasing the dependency on known labels and enabled the training on massive unlabeled data [18]. They also have shown the superior capacity in dealing with graphs which could thoroughly learn the discriminative representations of nodes [19, 20]. Meanwhile, graph contrastive learning (GCL) modules have already been widely used to handle the pairwise relationship prediction tasks among biology entities in the bioinformatics area. For example, SGCL-DTI first generated the topology and semantic graph for drug–target pairs and established a contrastive loss function to guide the learning process in a supervised manner to obtain embeddings of drugs and targets [21]. To predict protein–peptide binding residues, PepBCL established a novel contrastive learning strategy to learn the embeddings of binding residues based on the imbalanced dataset [22]. To predict cancer drug response problems, GraphCDR first constructed two different drug–cell line association networks and adopted the contrastive learning strategy to enhance its ability in learning the feature representations of nodes [23]. Besides, MIRACLE took multiview graph contrastive learning strategy to predict drug-drug interactions, which could capture molecule structure in the inter-view and interactions in the intra-view between molecules simultaneously [24]. To fully learn the embedding of nodes in the heterogeneous networks, HeCo generated network schema view and meta-path view based on HINs, and applied the cross-view contrastive mechanism to capture the information in local and high-order structures simultaneously [25]. In bioinformatics areas, generating different meaningful views appropriately is one essential step for these approaches above. Standard data augmentation approaches, such as node dropping or edge perturbation, are not trivial for common biological networks because they might damage the original graph structure and degrade the ability of prediction models in learning the feature representations [26, 27]. Meanwhile, as heterogeneous networks usually consist of multiple types of nodes and relations, GCL approaches should comprehensively mine the complex structure and rich semantics for learning the embeddings of nodes.

For the pairwise relationship prediction task, it is still a challenging problem to select the most informative negative samples from the candidate negative sample set [28]. Existing machine learning methods typically treat the known associations (labeled samples) between entities as the positive samples and the remained unconfirmed associations (unlabeled samples) as the candidate negative samples [29]. In this manner, there is an extreme imbalance between the number of positive and negative samples. What's more, with the negative under-sampling strategy, most approaches only randomly select a subset of negative samples from the whole candidate negative samples. [30]. For example, for the drug–target interaction prediction [31], miRNA–disease associations prediction [30, 32–34] and microbe–drug association prediction problems [13], these methods randomly selected the same number of negative samples as that of positive samples. A standard random under-sampling strategy often leads to the negligence of important and informative samples, and the introduction of meaningless and noisy samples [35]. Although some other models [36–38] improved the negative sampling strategy, they do not fully screen out the most informative negative samples that play an important role in the classifiers in the training process, which may largely limit their prediction capability.

Motivated by GCL approaches, we adopt the structure-enhanced contrastive learning strategy to obtain deep representations of microbes and drugs. Since microbes and drugs have multisource information, we first measure their respective similarity from different perspectives and construct the integrated similarity networks. Then to fully capture the complex structure and rich semantics of microbe–drug association network, we establish the meta-path-induced networks based on different meta-paths. Therefore, the similarity networks and meta-path-induced networks form the two views for contrastive learning. So we utilize the meta-path-induced networks of microbes and drugs to enhance their feature representations learned from the similarity networks. Besides, we adopt the self-paced negative sampling strategy to select the most informative negative samples, which aim to improve the capability of the prediction model.

In this study, we put forward a novel method that employs Structure-enhanced Contrastive learning and Self-paced negative sampling strategy to identify potential Microbe-Drug Associations (SCSMDA). Firstly, SCSMDA constructs the similarity networks of microbes and drugs, as well as their different meta-path-induced networks. Then, we employ the meta-path-induced networks of microbes and drugs to enhance their feature representations learned from the similarity networks with the contrastive learning strategy. After that, we utilize the self-paced negative sampling strategy to select the most informative negative samples to train the MLP classifier. Lastly, SCSMDA predicts the potential microbe–drug associations with the trained MLP classifier.

The workflow of SCSMDA is displayed in Figure 1. Our main contributions are summarized as follows:

1) SCSMDA constructs the similarity networks with the multisource information of microbes and drugs, and the meta-path-induced networks of microbes and drugs with different meta-paths.
2) SCSMDA employs the structure-enhanced contrastive learning strategy to obtain the discriminative embeddings of microbes and drugs in a self-supervised manner based on their similarity networks and meta-path-induced networks.
3) SCSMDA adopts the self-paced negative sampling strategy to select the most informative negative samples for training the MLP classifier.
4) Experimental results on three datasets indicate that SCSMDA outperforms other baseline approaches in microbe–drug association prediction tasks.

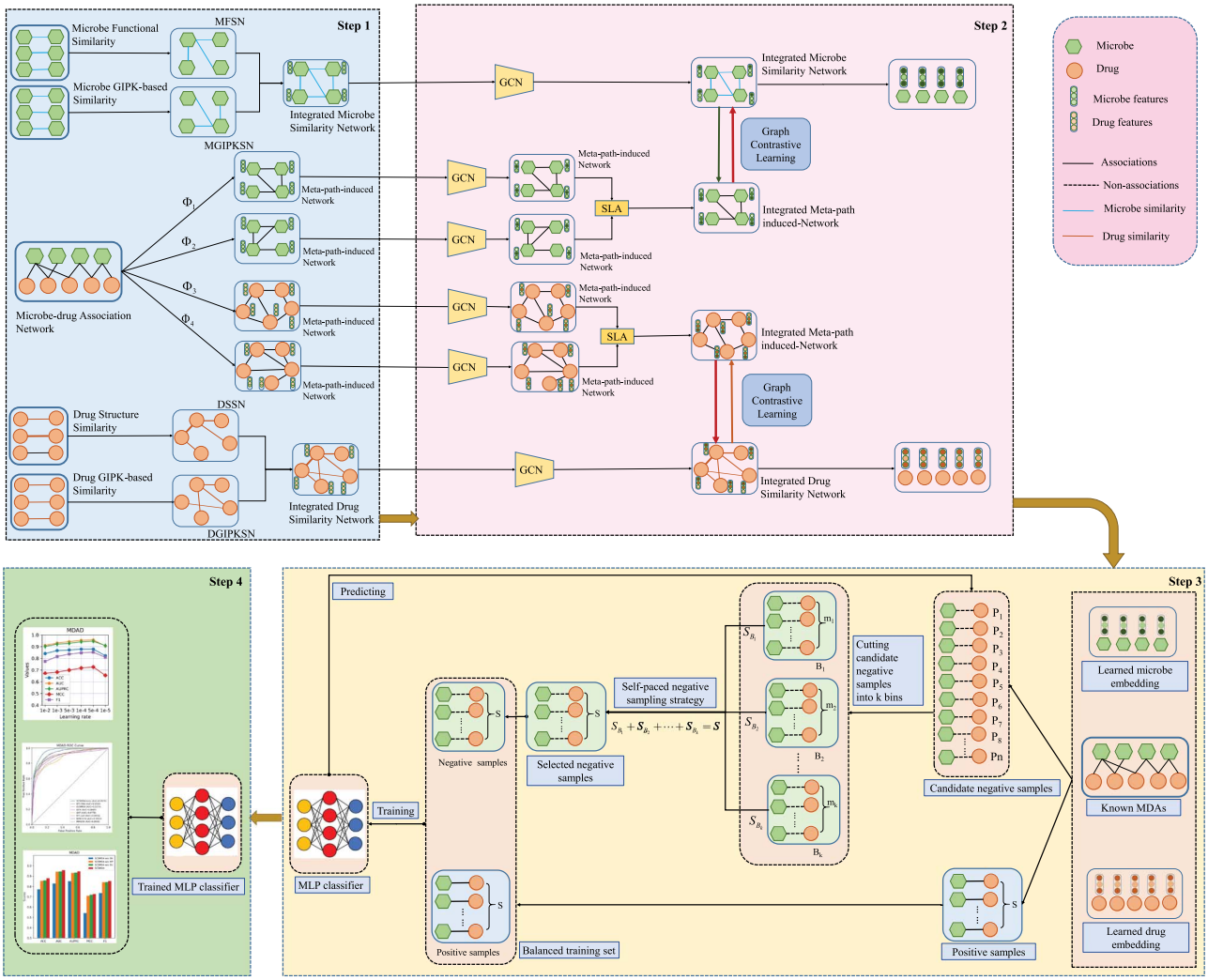

Figure 1. The overall workflow of SCSMDA. In step 1, SCSMDA constructs the similarity networks of microbes and drugs with their multisource information, as well as their different meta-path-induced networks. In step 2, we employ the meta-path-induced networks of microbes and drugs to enhance their feature representations learned from the similarity networks with the contrastive learning strategy. In step 3, SCSMDA adopts the self-paced negative sampling strategy to select the most informative negative samples for training the MLP classifier. In step 4, SCSMDA predicts the potential microbe–drug associations with the trained MLP classifier. In the figure, Φ_1, Φ_2, Φ_3 and Φ_4 denote meta-path MDM, MDMDM, DMD and DMDMD, respectively. SLA represents the semantic level attention.

Materials and methods

In this section, we will first briefly describe the experiment datasets and basic concepts used in SCSMDA. Then, the integrated similarity networks and meta-path-induced networks of microbes and drugs are established. Next, SCSMDA learns the embeddings of microbes and drugs with structure-enhanced contrastive learning strategy. After that, we utilize the self-paced negative sampling strategy to select the most informative negative samples and train the MLP classifier. Lastly, the loss function and some implementation details are presented.

Data collection

Currently, there are mainly three different known microbe–drug association datasets, which are MDAD [39], aBiofilm [40] and DrugVirus [41]. We collect these public datasets from the research [13] (https://github.com/longyahui/GCNMDA). Specifically, MDAD mainly contains 173 microbes and 1373 drugs involving 2470 associations. For aBiofilm dataset, it consists of 2884 microbe–drug associations between 140 microbes and 1720 drugs. For DrugVirus dataset, there are 95 microbes and 175 drugs including 933 microbe–drug associations between them. The statistics about these datasets are displayed in Table 2.

In each dataset, the association relationships between microbes and drugs can be established as one bipartite network. Without loss generality, the corresponding adjacency matrix can be denoted as $A \in \mathbb{R}^{N_m \times N_d}$, where N_m and N_d represent the number of microbes and drugs in the bipartite network. A_{ij} will be 1 if there is one association between m_i and d_j, and 0 otherwise.

Basic concept

Definition 1. Heterogeneous Information Network (HIN). One heterogeneous information network could be defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the entity type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{A}$ and a relation type mapping $\varphi : \mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V}$ and $\mathcal{A}$ denote the entity set and entity type set, and $\mathcal{E}$ and $\mathcal{R}$ denote the relation set and relation type set. Network $\mathcal{G}$ will be one homogeneous information network

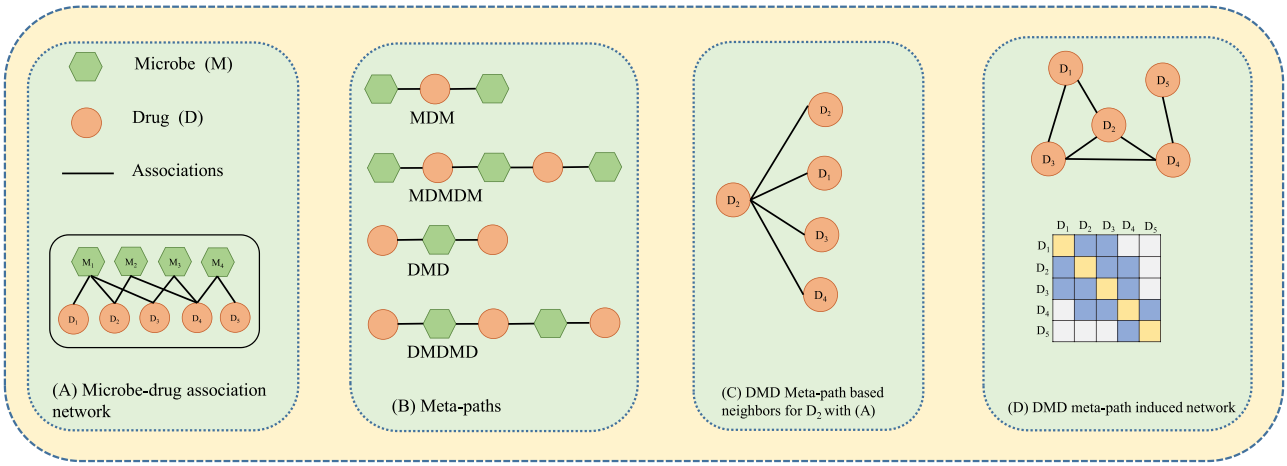

Figure 2. A toy example for SCSMDA. (A) Microbe–drug association network. (B) Four meta-paths involved in SCSMDA, which are MDM, MDMDM, DMD and DMDMD. (C) Drug D_2 and its DTD meta-path-based neighbors D_1, D_2, D_3 and D_4 based on the microbe–drug association network in (**A**). (D)The meta-path-induced network with DMD based on the network in **A**.

Table 1. Main notations in this research

Notations	Descriptions
$\mathcal{G}$	Heterogeneous Information Network
Φ	Meta-path
A	microbe–drug association matrix
A_Φ	Meta-path-induced matrix under Φ
h	Initial features of nodes
h'	Projected feature of nodes
sn	The integrated similarity network
mp	The integrated meta-path-induced network
$z^{sn}_{m_i}$	The embedding of microbe m_i learned from sn
$z^{mp}_{m_i}$	The embedding of microbe m_i learned from mp
$z^{sn}_{d_j}$	The embedding of drug d_j learned from sn
$z^{mp}_{d_j}$	The embedding of drug d_j learned from mp
z_{m_i}	The final embedding of microbe m_i
z_{d_j}	The final embedding of drug d_j
$\mathcal{H}$	The Hardness function
$\mathcal{N}^{\Phi}_{v_i}$	Meta-path-based neighbors for v_i with Φ
(i,j)	The node pair of microbe m_i and drug d_j
y_{ij}	The ground truth of the node pair (i,j)
$\hat{y}_{ij}$	The predicted score of the node pair (i,j)
Y^+	Positive MDAs in the training set
Y^-	Selected negative MDAs in the training set
MLP	The Multilayer Perceptron

Table 2. The statistics for microbe–drug association datasets.

Datasets	# Microbes	# Drugs	# Associations
MDAD[39]	173	1,373	2,470
aBiofilm[40]	140	1,720	2,884
DrugVirus[41]	95	175	933

if $|\mathcal{A}| + |\mathcal{R}| = 2$. Otherwise, it will be one heterogeneous information network.

Example. The microbe–drug association network (Figure 2A) could be treated as one HIN, since there are two types of nodes which are microbe and drug, and one type of link, which is the association relationship.

Definition 2. Meta-paths. Generally, one meta-path Φ with l nodes can be defined as $N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} N_l$, which is abbreviated as $N_1N_2\cdots N_l$. The composition relation between node N_1 and N_l is formulated as $R = R_1 \circ R_2 \circ \cdots \circ R_l$, where $\circ$ is the composition operator on relations.

Example. In the microbe–drug HIN (Figure 2A), two drugs can be connected by different meta-paths (Figure 2B), such as DMD and DMDMD. This type of meta-paths usually has a certain biological meaning. For example, DMD indicates that if two drugs interact with one common microbe, they should have a higher similarity with consistent functionality.

Definition 3. Meta-path-based neighbors. Suppose there is one node named v_i and one meta-path Φ, its meta-path-based neighbors $\mathcal{N}^{\Phi}_{v_i}$ can be defined as the nodes that connect with v_i according to the meta-path Φ.

Example. As is shown in Figure 2C, for drug D_1, its DMD meta-path-based neighbors are D_1, D_2, D_3 and D_4 based on Figure 2C.

Microbe and drug similarity network construction

Microbe similarity network construction

SCSMDA measures the similarity of microbes from two aspects. The 1st kind of similarity is called the microbe functional similarity. Suppose there are two microbes named m_i and m_j respectively, their microbe functional similarity can be denoted as $FM(m_i, m_j)$. SCSMDA measures the microbe functional similarity between all microbe pairs and finally establishes the Microbe Functional Similarity Network. The detailed calculation process is presented by Kamneva [42] and Long [13].

The 2nd type of microbe similarity is called the Gaussian-interaction-profile-kernel-based similarity. The basic assumption for this type of similarity is that similar microbes (drugs) interacting with similar drugs (microbes) will have similar profiles. Specifically, suppose there is one microbe–drug association

matrix named A, the interaction profiles for microbe m_i and m_j can be denoted as the i-th and j-th row in association matrix A, which are represented as $A(m_i)$ and $A(m_j)$. So, the Gaussian interaction profile kernel-based similarity for microbe m_i and m_j is formulated as:

$$GM(m_i, m_j) = exp\left(-\eta_m \|A(m_i) - A(m_j)\|^2\right), \quad (1)$$

where η_m is the normalized kernel bandwidth, which is calculated as:

$$\eta_m = \eta'_m / \left(\frac{1}{N_m}\sum_{i=1}^{N_m}\|A(m_i)\|\right), \quad (2)$$

where η'_m is always set to 1. SCSMDA measures all the similarities of all microbe pairs and constructs the Microbe Gaussian-Interaction-Profile-Kernel-based Similarity Network.

Suppose there are two microbes named m_i and m_j, and their functional similarity and Gaussian-interaction-profile-kernel-based similarity are $FM(m_i, m_j)$ and $GM(m_i, m_j)$, the integrated microbe similarity S_m is defined as:

$$S_m(m_i, m_j) = \begin{cases} \frac{FM(m_i,m_j)+GM(m_i,m_j)}{2} & if\ FM(m_i, m_j) \neq 0 \\ GM(m_i, m_j) & otherwise \end{cases} \quad (3)$$

SCSMDA measures the integrated similarities for all the microbe pairs and then constructs the integrated microbe similarity network.

Drug similarity network construction

Meanwhile, we also measure the similarity of drugs from two aspects. The 1st one is the drug structure-based similarity proposed by Hattori [43]. For two drugs named d_i and d_j, their structure-based similarity can be represented as $DS(d_i, d_j)$. After calculating all the similarities between all drug pairs, we can establish the Drug Structure-based Similarity Network.

The 2nd similarity between drugs is the Gaussian-interaction-profile-kernel-based similarity. Similar to the Gaussian-interaction-profile-kernel-based similarity of microbes, the drug Gaussian-interaction-profile-kernel-based similarity between d_i and d_j can be defined as:

$$GD(d_i, d_j) = exp\left(-\eta_d \|A(d_i) - A(d_j)\|^2\right), \quad (4)$$

where $A(d_i)$ and $A(d_j)$ represent the interaction profiles, which are defined as the i-th and j-th columns in microbe–drug association matrix A. And η_d is the normalized kernel bandwidth, which is calculated as:

$$\eta_d = \eta'_d / \left(\frac{1}{N_d}\sum_{i=1}^{N_d}\|A(d_i)\|\right) \quad (5)$$

where η'_d is always set to 1. SCSMDA measures the similarity of all drug pairs and constructs the Drug Gaussian-Interaction-Profile-Kernel-based Similarity Network.

For two microbes named d_i and d_j and their drug structure-based similarity and drug Gaussian-interaction-profile-kernel-based similarity are $DS(d_i, d_j)$ and $GD(d_i, d_j)$ respectively, the integrated microbe similarity S_d is defined as:

$$S_d(d_i, d_j) = \begin{cases} \frac{DS(d_i,d_j)+GD(d_i,d_j)}{2} & if\ DS(d_i, d_j) \neq 0 \\ GD(d_i, d_j) & otherwise \end{cases} \quad (6)$$

SCSMDA measures the integrated similarities for all the drug pairs and then constructs the integrated drug similarity network.

Meta-path-induced network construction

The microbe–drug association network can be regarded as one HIN with complex structure and rich semantics. Meta-paths could comprehensively reflect the structure of HINs and have been widely employed to capture rich semantic meanings in HINs. Therefore, SCSMDA establishes different meta-path-induced networks for microbes and drugs according to their diverse meta-paths.

In this study, SCSMDA mainly adopts two meta-paths named $\Phi_1 = MDM$ and $\Phi_2 = MDMDM$ for microbes, and two meta-paths named $\Phi_3 = DMD$ and $\Phi_4 = DMDMD$ for drugs to establish their corresponding meta-path-induced networks. For the microbe–drug association network represented as A, given meta-path $\Phi_1 = MDM$ and $\Phi_2 = MDMDM$, the corresponding meta-path-induced networks for microbes can be formulated as:

$$A_{\Phi_1} = A \times A^T \quad (7)$$

$$A_{\Phi_2} = A^2_{\Phi_1} = (A \times A^T)^2. \quad (8)$$

Meanwhile, the meta-path-induced networks for drugs with Φ_3 and Φ_4 can be represented as

$$A_{\Phi_3} = A^T \times A \quad (9)$$

$$A_{\Phi_4} = A^2_{\Phi_3} = (A^T \times A)^2. \quad (10)$$

A toy example for constructing the meta-path-induced network has been represented in Figure 2D.

Node feature transformation

Since there are two different types of nodes in microbe–drug association network and their initial features belong to different spaces, we need to transform their features into one common vector space. Without loss generality, for one node v_i with type ϕ_{v_i}, SCSMDA maps its initial features into one shared space denoted as:

$$h'_{v_i} = \sigma(W_{\phi_{v_i}} \cdot h_{v_i} + b_{\phi_{v_i}}), \quad (11)$$

where $h'_{v_i} \in \mathbb{R}^{d\times 1}$ is the projected feature for node v_i, $\sigma(\cdot)$ is the activation function, $W_{\phi_{v_i}}$ is the type-specific mapping matrix and b_ϕ is the vector bias.

Embeddings learning from the integrated similarity networks

Particularly, GCNs have exhibited the great expressive ability in learning the embeddings of nodes in graphs [16]. For *vanilla* GCN

[44], one-layer graph convolution encoder on graph G with a symmetric adjacency matrix Q can be represented as:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{Q}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}), \tag{12}$$

where $\sigma(\cdot)$ is the activation function. $\widetilde{Q} = Q + I$ and I is the identity matrix with the same shape as Q, $\widetilde{D}$ is the degree matrix of $\widetilde{Q}$, $W^{(l)}$ is the learnable weights at l^{th} layer, $H^{(l)}$ denotes the representations of nodes at l^{th} layer. The output representations of nodes at l^{th} layer can be input into the next layer of GCNs. In this way, we can get the nodes embeddings at any layer.

Suppose the integrated microbe similarity network is S_m, the embedding of microbes at $l+1$ layer can be formulated as follows:

$$H_{S_m}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{S}_m\widetilde{D}^{-\frac{1}{2}}H_{S_m}^{(l)}W_{S_m}^{(l)}), \tag{13}$$

where $\widetilde{S}_m = S_m + I$ and I is the identity matrix with the same shape as S_m, $\widetilde{D}$ is the degree matrix of $\widetilde{S}_m$, $H_{S_m}^{(l)}$ denotes the representations of microbes at l^{th} layer.

Similarly, suppose the integrated drug similarity network is S_d, the embedding of drugs at $l+1$ layer can be formulated as:

$$H_{S_d}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{S}_d\widetilde{D}^{-\frac{1}{2}}H_{S_d}^{(l)}W_{S_d}^{(l)}), \tag{14}$$

where $\widetilde{S}_d = S_d + I$ and I is the identity matrix with the same shape as S_d, $\widetilde{D}$ is the degree matrix of $\widetilde{S}_d$, $H_{S_d}^{(l)}$ denotes the representations of microbes at l^{th} layer.

The ultimate embeddings for microbes m_i and drugs d_j learned from the integrated similarity networks can be represented as $z_{m_i}^{sn}$ and $z_{d_j}^{sn}$.

Embedding learning with meta-path-induced networks

SCSMDA generates two different meta-path-induced networks for microbes and drugs, respectively. Since microbes and drugs have similar learning module structures with meta-path-induced networks, we only take microbes as an example to show the process that SCSMDA learns their embeddings with *vanilla* GCNs.

The meta-path-induced network for microbes with Φ_n is denoted as A_{Φ_n}, where $n \in \{1, 2\}$. We apply *vanilla* GCNs on A_{Φ_n} to learn the embeddings of microbes, which can be formulated as:

$$H_{A_{\Phi_n}}^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}_{\Phi_n}\widetilde{D}^{-\frac{1}{2}}A_{\Phi_n}^{(l)}W_{A_{\Phi_n}}^{(l)}\right), n \in \{1, 2\} \tag{15}$$

where $H_{A_{\Phi_n}}^{(l+1)}$ is the embeddings of microbes at l-th layer.

SCSMDA adopts the Semantic Level Attention (SLA) to obtain the final embeddings of microbes from different meta-path-induced networks. Suppose there is one microbe m_i, its embedding learned from A_{Φ_n} is represented as $h_{m_i}^{A_{\Phi_n}}$, where $n \in \{1, 2\}$. The final embedding from the meta-path-induced networks for m_i is denoted as:

$$z_{m_i}^{mp} = \sum_{n=1}^{N} \beta_{\Phi_n} \cdot h_{m_i}^{A_{\Phi_n}}, \tag{16}$$

where β_{Φ_n} is the learned weight for meta-path Φ_n and can be calculated as

$$w_{\Phi_n} = \frac{1}{|V|}\sum_{m_i \in V} \boldsymbol{a}_{mp}^T \cdot tanh\left(\boldsymbol{W}_{mp} \cdot h_{m_i}^{\Phi_n} + \boldsymbol{b}_{mp}\right) \tag{17}$$

$$\beta_{\Phi_n} = \frac{exp(w_{\Phi_n})}{\sum_{n=1}^{N} exp\left(w_{\Phi_n}\right)}, \tag{18}$$

where $W_{mp} \in \mathbb{R}^{d \times d}$ and $b_{mp} \in \mathbb{R}^{d \times 1}$ are the two learnable parameters, and $\boldsymbol{a}_{mp}$ represents the semantic-level attention vector.

Similarly, suppose there is one drug d_j, its final embeddings learned from the meta-path-induced network A_{Φ_n} where $n \in \{3, 4\}$ can be represented as $z_{d_j}^{mp}$.

Structure-enhanced contrastive learning strategy

After getting two types of embeddings $z_{m_i}^{sn}$ and $z_{m_i}^{mp}$ for microbe m_i, we feed them into one MLP layer and get the embeddings used for calculating the contrasting loss:

$$z_{m_i}^{sn}\_proj = W^{(2)}\sigma\left(W^{(1)}z_{m_i}^{sn} + b^{(1)}\right) + b^{(2)}, \tag{19}$$

$$z_{m_i}^{mp}\_proj = W^{(2)}\sigma\left(W^{(1)}z_{m_i}^{mp} + b^{(1)}\right) + b^{(2)}, \tag{20}$$

where σ is the ReLU nonlinear function. The parameters $W^{(1)}$,$W^{(2)}$,$b^{(1)}$ and $b^{(2)}$ are shared by the two embeddings learned from the similarity networks and meta-path-induced networks.

Generally, traditional contrastive learning approaches only treat the same instances at different augmented views as the positive sample and treat the different instances as negative samples [45]. Differently, SCSMDA chooses a novel positive selection strategy that if two microbes are connected by enough meta-paths, they could also be regarded as positive samples. Specifically, for two microbes m_i and m_j, SCSMDA first counts the meta-paths connecting two microbes, which can be formulated as:

$$\mathbb{C}_{m_i}(m_j) = \sum_{n=1}^{N} \mathbb{I}\left(m_j \in \mathcal{N}_{m_i}^{\Phi_n}\right), \tag{21}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathcal{N}_{m_i}^{\Phi_n}$ is the meta-path neighbors of m_i under Φ_n. Then, we construct set $S_{m_i} = \{m_j | m_j \in V \text{ and } \mathbb{C}_{m_i}(m_j) \neq 0\}$ and sort the elements in S_{m_i} with the value of $\mathbb{C}_{m_i}(\cdot)$ in a descending order. After that, we define a threshold (named T_{pos}) and select the top T_{pos} microbes from S_{m_i}. These selected microbes form the conditional positive sample set denoted as $\mathbb{P}_{m_i}$. In particular, we treat the same microbe in $\mathbb{P}_{m_i}$ as the instinct positive pair. The remained microbes in S_{m_i} are treated as conditional negative samples denoted as $\mathbb{N}_{m_i}$. A toy example for instinct positive pair, conditional positive pairs and conditional negative pairs is displayed in Figure 3.

Based on the conditional positive sample set $\mathbb{P}_{m_i}$ and the conditional negative sample set $\mathbb{N}_{m_i}$, the contrastive loss from the integrated similarity network can be defined as:

$$\mathcal{L}_{m_i}^{sn} = -log\frac{\sum_{m_j \in \mathbb{P}_{m_i}} exp\left(sim\left(z_{m_i}^{sn}\_proj, z_{m_j}^{mp}\_proj\right)/\tau\right)}{\sum_{m_k \in \mathbb{P}_{m_i} \cup \mathbb{N}_{m_i}} exp\left(sim\left(z_{m_i}^{sn}\_proj, z_{m_k}^{mp}\_proj\right)/\tau\right)}, \tag{22}$$

where $sim(z_{m_i}^{sn}, z_{m_j}^{mp})$ is the cosine similarity between microbe m_i and m_j, and τ is the temperature parameter.

Meanwhile, the contrastive learning loss in the integrated meta-path-induced network $\mathcal{L}_{m_i}^{mp}$ is similar to $\mathcal{L}_{m_i}^{sn}$, which can be formulated as:

$$\mathcal{L}_{m_i}^{mp} = -log\frac{\sum_{m_j \in \mathbb{P}_{m_i}} exp\left(sim\left(z_{m_i}^{mp}\_proj, z_{m_j}^{sn}\_proj\right)/\tau\right)}{\sum_{m_k \in \mathbb{P}_{m_i} \cup \mathbb{N}_{m_i}} exp\left(sim\left(z_{m_i}^{mp}\_proj, z_{m_k}^{sn}\_proj\right)/\tau\right)}. \tag{23}$$

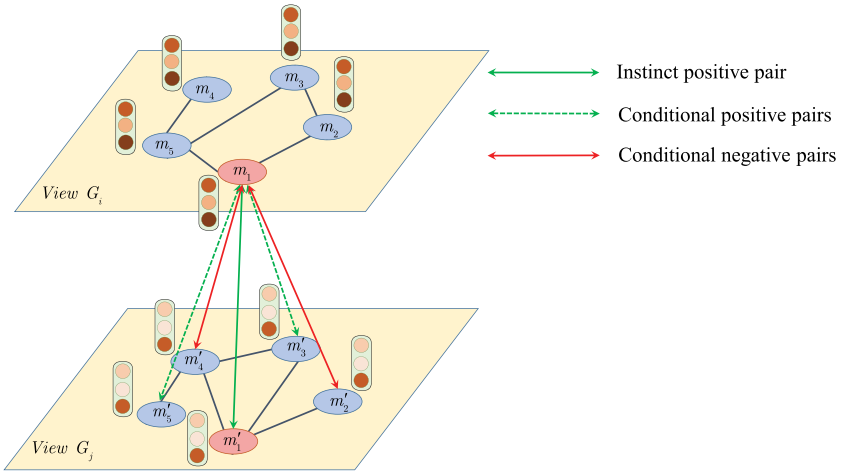

Figure 3. The positive pair selecting strategy of SCSMDA. G_i and G_j are two different views. m_1 and m'_1 are the same nodes at G_i and G_j and (m_1, m'_1) is the instinct positive pair. m'_2 and m'_5 will be the conditional positive samples for m_1 if they are connected by enough meta-paths and (m_1,m'_2) and (m_1,m'_5) could be the conditional positive pairs. Meanwhile, (m_1, m'_3) and (m_1,m'_4) are the conditional negative pairs if (m_1, m'_3) and (m_1,m'_4) don't connect by enough meta-paths.

The overall loss function for learning the embeddings of microbes can be defined as:

$$\mathcal{J}_m = \frac{1}{|V|}\sum_{m_i \in V}\left[\lambda_m \cdot \mathcal{L}^{sn}_{m_i} + (1-\lambda_m)\cdot \mathcal{L}^{mp}_{m_i}\right], \tag{24}$$

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

SCSMDA learns the embeddings of nodes from the integrated similarity network by aggregating Information from their direct neighbors, making it could capture the local structure. Meanwhile, SCSMDA could also learn embedding from the meta-path-induced networks with multiple meta-paths, aiming at capturing the high-order structure Information. In our study, the proposed structure-enhanced contrastive strategy employs the representations of microbes and drugs from meta-path-induced networks to enhance their embeddings learned from the similarity networks with the contrastive learning strategy. SCSMDA adopts the embeddings of microbes learned from the integrated similarity network as their final embedding.

Meanwhile, SCSMDA learns the embeddings of drugs in a similar way, and its overall loss objection is defined as:

$$\mathcal{J}_d = \frac{1}{|V|}\sum_{d_j \in V}\left[\lambda_d \cdot \mathcal{L}^{sn}_{d_j} + (1-\lambda_d)\cdot \mathcal{L}^{mp}_{d_j}\right], \tag{25}$$

where parameter λ_d is the coefficient to balance the contributions of the similarity network and meta-path-induced network. We could optimize SCSMDA via backpropagation for learning the feature representations of microbes and drugs. Lastly, the representations from the integrated similarity networks for microbes and drugs are regarded as the final embeddings denoted as z_{m_i} and z_{d_j}.

Self-paced negative sampling strategy

In the microbe–drug association datasets, all the known microbe–drug associations form the positive sample set denoted as P, whereas all the remaining microbe–drug associations are regarded as the candidate negative samples denoted as N. The number of positive and negative samples in this study has a relationship that $|N| \gg |P|$.

Previous research always randomly selects the same number of negative samples with that of the positive samples from the candidate negative sample set, which does not fully consider the specificity of negative samples. Selecting the most informative samples from the candidate negative sample set is a challenging task, which affects the capability of the prediction model. Here we employ the self-paced negative-sampling strategy motived by SPE [46] to choose the most informative negative samples.

The self-paced negative sampling strategy divides samples in N into three classes with the Hardness function $\mathcal{H}$, which are trivial samples, noise samples and borderline samples. The trivial samples are scored with small values by $\mathcal{H}$ indicating that they are well classified by the classifier, whereas the noise samples are scored with large values by $\mathcal{H}$ meaning that they may be false negative samples. These two types of samples should be selected as the negative samples with smaller probabilities for training the classifier. Correspondingly, we should focus on the borderline samples with scores around 0.5, since these samples are the most informative and should be selected as the negative samples with larger probabilities for training the classifier.

There are four steps for self-paced negative sampling strategy in SCSMDA, which are listed below.

- Step one: SCSMDA predicts the values for all the candidate negative microbe–drug association pairs with the MLP classifier $f(\cdot)$.
- Step two: SCSMDA cuts all the candidate negative samples into k bins with respect to values scored by the hardness function $\mathcal{H}$, which can be formulated as:

$$B_l = \{(x,y)|\frac{l-1}{k} \le \mathcal{H}(x,y,f) < \frac{l}{k}\}, \mathcal{H} \in [0,1], \tag{26}$$

where k is a hyper-parameter. B_l is the negative sample set for l-th bin where $l \in \{1,2,\ldots,k\}$. The hardness function used in SCSMDA is defined as:

$$\mathcal{H}(x,y,f) = |f(x) - y|, \tag{27}$$

where $f(x)$ represents the MLP classifier's output probability score of sample x and y is the ground-truth label of sample x.

- Step three: SCSMDA employs the self-paced negative strategy to select the negative samples from k bins and obtains the negative sample set, which can be denoted as:

$$N_0 = \{x_{lj} | l \in [1, k], j \in [1, S_{B_l}]; l, j \in \mathbb{N}^+\}, \tag{28}$$

where k is the number of bins, S_{B_l} is the number of negative samples selected from l-th bin, and x_{ln} denotes the j-th selected sample from l-th bin B_l. Parameter S_{B_l} is defined as:

$$\begin{cases} S_{B_l} = \frac{w_l}{\sum_t w_t} \cdot |P|, t = 1, \cdots, k \\ w_l = \frac{1}{h_l + \alpha}, l = 1, \cdots, k \\ \alpha = tan(\frac{i\pi}{2n}) \\ h_l = \sum_{x_{lj} \in B_l} \mathcal{H}(x_{lj}, y_{lj}, f)/|B_l|, l = 1, \cdots, k, \end{cases}$$

where w_l represents the normalized sampling weight of l-th bin, α is called the self-paced factor and h_l denotes the average hardness contribution for l-th bin. Besides, i denotes iteration number.

- Step four: The selected negative samples N_0 and all the known positive samples P are composed of the training set to train the MLP classifier and begin the next iteration.

The algorithm for the self-paced negative sampling strategy is shown in Algorithm 1.

Algorithm 1 : The self-paced negative sampling strategy

Input:
Hardness function $\mathcal{H}$
Classifier to be trained f
Positive sample set P
Candidate negative sample set N
Number of bins k
Number of training epochs n

Train:
Train f with negative samples N_0 and positive samples P, where $|N_0| = |P|$ and N_0 is selected from N randomly;

for $i = 1$ to n **do**
  Predict association scores for all samples in N with f ;
  Cut all the candidate negative samples into k bins denoted as $B_1, \cdots, B_k$ via Eq.26 and Eq.27;
  Calculate the average hardness value for each bin: $h_l = \frac{\sum_{j=1}^{|B_l|} f(x_{lj})}{|B_l|}, l = 1, 2, \ldots, k$, and $x_{lj} \in B_l$;
  Update self-paced factor $\alpha = tan(\frac{i\pi}{2n})$;
  Calculate the weight of each bin: $w_l = \frac{1}{h_l + \alpha}$;
  Calculate the under-sampling number for each bin: $S_{B_l} = \frac{w_l * |P|}{\sum_{j=1}^{k} w_j}, l = 1, 2, \ldots k$;
  $N_0 = \varnothing$;
  for $l = 1$ to k **do**
    Randomly select S_{B_l} negative samples from B_l and add to N_0;
  end for
  Train f using new negative sample set N_0 and positive sample set P;
  Calculate model losses by Cross-Entropy loss function;
  Update model parameters via backpropagation, get f;
end for
Return: Trained model f

Final decoder

In this research, we adopt MLP as the final decoder, which first employs the embeddings of microbes and drugs as its input and then performs the element-wise multiplication operation on the embeddings of microbes and drugs. Lastly, the association probability score $\hat{y}_{ij}$ for microbe m_i and drug d_j can be formulated as:

$$\hat{y}_{ij} = Sigmoid(Q_1(ReLU(Q_2(z_{m_i} \odot z_{d_j})))) \tag{29}$$

where z_{m_i} and z_{d_j} denote the embeddings for microbe m_i and drug d_j. The operation $\odot$ denotes the element-wise multiplication for microbe $z_{m_i} \in R^{F'}$ and drugs $z_{d_j} \in R^{F'}$, and $Q_1 \in R^{1 \times F'}$ and $Q_2 \in R^{F' \times F'}$ are the learnable matrices. Besides, *ReLU* and *Sigmoid* are the two activation functions.

Loss function

SCSMDA applied the binary cross-entropy as the loss function in microbe–drug association prediction problem because of its effective performance on the binary-classification task. The binary cross-entropy loss (denoted as LB) used in SCSMDA is defined as

$$\mathcal{L}_B = - \sum_{(i,j) \in Y^+ \cup Y^-} y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) log(1 - \hat{y}_{ij}), \tag{30}$$

where (i, j) denotes the microbe–drug sample for microbe m_i and drug d_j, Y^+ and Y^- are positive and negative microbe–drug sample subsets for training, respectively. If one microbe–drug pair $(i, j) \in Y^+$, the ground truth y_{ij} is 1. If $(i, j) \in Y^-$, the ground truth y_{ij} is 0. The prediction association value is represented as $\hat{y}_{ij}$.

Meanwhile, coupled with the two loss functions at the structure-enhanced contrastive learning strategy (Eq. 24 and Eq. 25), the final overall loss function $\mathcal{L}$ for SCSMDA is formulated as:

$$\mathcal{L} = \mathcal{L}_B + \mathcal{J}_m + \mathcal{J}_d. \tag{31}$$

Implementation details

SCSMDA initializes the learnable parameters with Glorot initialization [47] and trains the model with Adam [48]. We adopt the grid search strategy to tune parameters for SCSMDA. Specifically, the learning rate is set to 0.0005 and τ is tuned to 0.5. The final embedding sizes of drugs and microbes are both 128. The numbers of GCN layers and MLP layers are equal to 1. The best number of positive pairs k is 10. Besides, during the training process, the dropout values for the encoder on the integrated similarity network and meta-path-induced network are 0.95 and 0.3, respectively, and SCSMDA achieves the highest evaluation values when the number of epochs is 1000.

Besides, we implement our mode using a software environment with PyCharm Community Edition 2022.1.1 version and libraries with Python v3.9.5, Pytorch v1.11.0, NumPy v1.22.3, sci-kit-learn v1.1.1, scipy v1.9.3, and tqdm v4.64.0. All experiments were performed on hardware with a desktop computer with one Intel(R) Core(TM) i5-12600KF CPU and one NVIDIA RTX3060 8GB GPU. The detailed Implementation information has been published on GitHub (https://github.com/Yue-Yuu/SCSMDA-master).

Time complexity analysis

As shown in Figure 1, there are mainly three steps for training SCSMDA, which are the construction of similarity and meta-path-induced networks, and the structure-enhanced contrastive learning strategy, the self-paced negative sampling strategy. Therefore, we analyze the time complexity for them one by one.

In step one, suppose there are m microbes and n drugs, SCSMDA first measures the similarity between microbes or drugs and their time complexity is $O(m^2/2) + O(n^2/2)$. For establishing the integrated microbe and drug similarity network, the time complexity is $O(m^2) + O(n^2)$. Besides, for establishing the meta-path-induced networks, their time complexities are $O(m^2n)$, (m^3), $O(n^2m)$, and $O(n^3)$ under meta-path Φ_1, Φ_2, Φ_3, and Φ_4, respectively. As a result, the total time complexity in this step is $O(m^2/2)+O(n^2/2)+O(m^2)+O(n^2)+O(m^2n)+O(n^2m)+O(m^3)+O(n^3) = O(m^3)+O(n^3)$.

In step two, SCSMDA adopts the GCNs to learn the embeddings of microbes and drugs. Suppose the layer number of GCNs is 1, and the initial and output feature dimensions of nodes are C and F, the time complexity for learning embeddings is $O(|E|CF)$, where E is the edge set of the input network to GCNs. Besides, since SCSMDA measure similarity between all the nodes for the contrastive learning strategy process, the time complexity is $O(m^2) + O(n^2)$. Therefore, the total time complexity in this step is $O(|E|CF) + O(m^2) + O(n^2)$.

In step three, SCSMDA selects the most informative samples from the candidate negative sample set. The positive sample set is denoted as P. Therefore, the number of the positive microbe–drug pairs is equal to $|P|$, and the number of the candidate negative microbe–drug pairs will be $(mn - |P|)$. The time complexity for performing one epoch is $O(mn - |P|)$. Suppose the epoch number is T, then the time complexity is $O(T(mn - |P|))$. Since $((mn) \gg |P|)$, the total time complexity in this step is $O(Tmn)$.

In summary, the total time complexity for training SCSMDA is the sum in these three steps above, which could be formulated as $O(m^3)+O(n^3)+O(|E|CF)+O(n^2)+O(m^2)+O(Tmn) = O(m^3)+O(n^3)+O(|E|CF) + O(Tmn)$. Since parameters C and F are constant, so the ultimate time complexity is $O(m^3) + O(n^3) + O(Tmn)$.

Results

In this section, we first describe the evaluation metrics widely used in our study. Then, a comprehensive comparison between SCSMDA and other baseline approaches will be presented from different aspects. After that, ablation study and parameter sensitivity analysis experiments for SCSMDA are extensively investigated. Lastly, we conduct case studies for two interested drugs.

Experimental setup and evaluation metrics

In this study, we adopt the 5-fold cross-validation (5-CV) strategy [49, 50] to evaluate the performance of SCSMDA as well as the baseline approaches on MDAD, aBiofilm and DrugVirus datasets, respectively. Specifically, for each dataset, all the known microbe–drug association pairs are treated as the positive samples and form the positive sample set, whereas all the remained unknown microbe–drug association pairs are treated as the candidate negative samples and form the candidate negative sample set. SCSMDA selects the same number of negative samples with that of positive samples according to the self-paced negative strategy from the candidate negative sample set. The positive samples and selected negative samples are constructed as the experimental dataset, and we conduct the 5-CV evaluation experiment on it.

For the 5-CV experiment, SCSMDA first divides the experimental dataset into five subsets with equal numbers. Then, each subset is treated as the test subset in turn and the remaining four subsets will be training subsets. In this way, we could calculate true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), respectively.

In addition, we mainly employ five metrics, which are area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPRC), accuracy (ACC), Matthews correlation coefficient (MCC) and F1 score to evaluate the performance of the SCSMDA as well as comparison methods. These five evaluation metrics are widely used in previous studies [30], and here we don't repeat them anymore.

To minimize the bias of the 5-CV strategy result, we perform the experiment five times for each method and then obtain the mean and standard deviation values of the scores.

Comparison with other baseline methods on AUC and AUPRC metrics

Here, we choose eight competitive approaches for comparison. These approaches are GCN [44], GAT [51], DTIGAT [52], NIMCGCN [53], MMGCN [54], GCNMDA[13], DTI-CNN [55] and Graph2MDA [16].

- GCN [44] is a semi-supervised learning approach. Here we feed microbe similarity network and drug similarity network into GCNs and learn their embeddings for predicting the association relationships.
- GAT [51] is one of the graph neural networks with the attention mechanism. We feed microbes similarity network and drug similarity network into GATs and obtain their feature representations for completing the microbe–drug association prediction tasks.
- DTIGAT [52] is originally employed to predict the interactions between proteins and drugs with the attention mechanism. Here we feed the microbe–drug association network into this model to learn the features of microbes and drugs.
- NIMCGCN [53] firstly adopts the GCNs to obtain the latent embeddings of miRNA and disease from their similarity networks and predicts miRNA–disease associations. We feed the microbe–drug association network into the model to predict microbe–drug associations.
- MMGCN [54] employs GCN encoder to obtain the embeddings of miRNA and disease in different similarity views and enhance the learned representations by utilizing multichannel attention mechanism.
- GCNMDA [13] builds a heterogeneous network for drugs and microbes, and then employs the GCN-based framework with conditional Random Field as well as attention mechanism techniques to discover microbe–drug associations, named GCNMDA.
- DTI-CNN [55] extracts the embeddings of drugs and proteins based on the heterogeneous networks and constructs a convolutional neural network model to infer their interactions with learned features from a denoising autoencoder model.
- Graph2MDA [16] adopts the variational graph autoencoder for learning the latent representations of microbes and drugs based on the multimodal attributed graphs and predicts MDAs with the deep neural network model.

We compare SCSMDA with other baseline methods on AUC and AUPRC metrics, and the corresponding results on MDAD, aBiofilm and DrugVirus datasets are shown in Figure 4. The proposed method SCSMDA achieves the best performance in all the SOTA

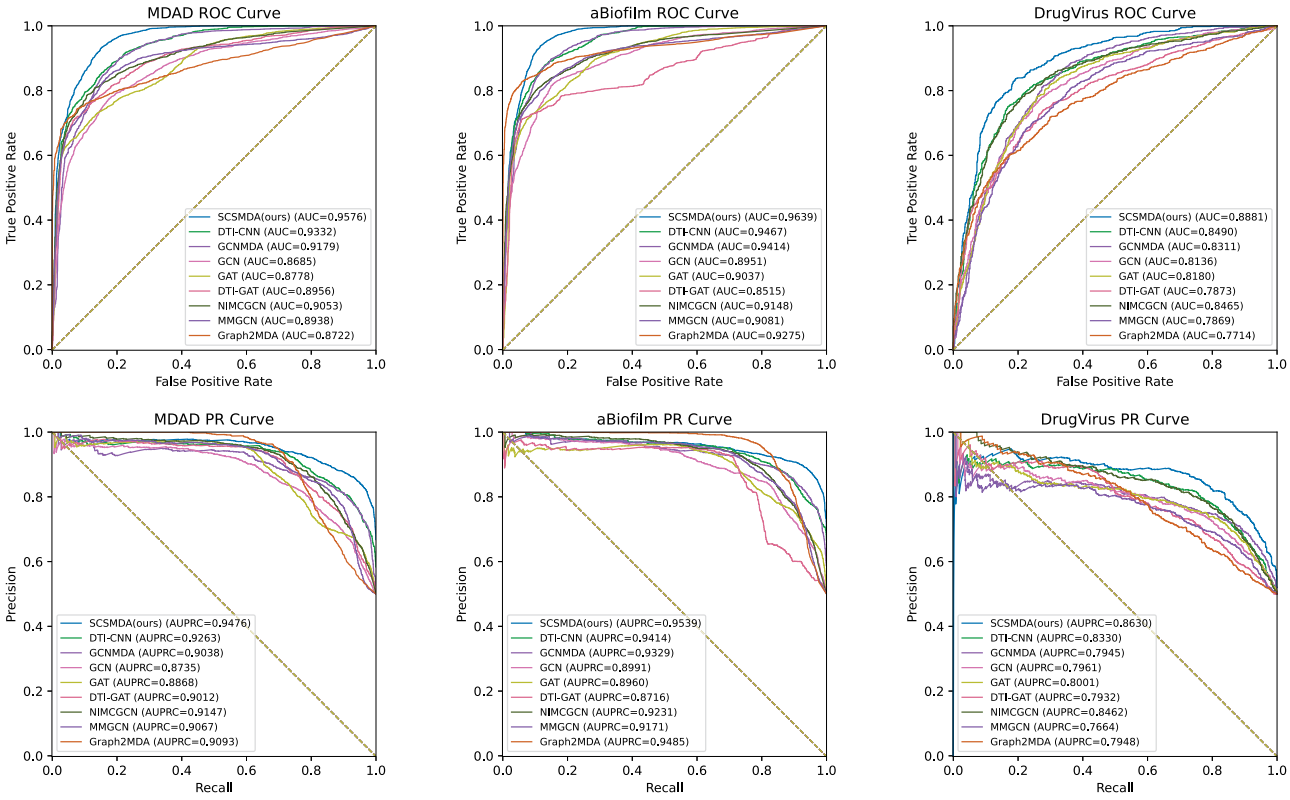

Figure 4. The ROC and PR curves of SCSMDA as well as the baseline methods for predicting microbe–drug associations on MDAD, aBiofilm and DrugVirus datasets.

approaches. In particular, the AUC values of SCSMDA on MDAD, aBiofilm and DrugVirus datasets are 0.9576, 0.9639 and 0.8881 respectively, whereas the AUPRC values of SCSMDA on these datasets are 0.9476, 0.9539 and 0.8630, respectively.

Besides, DTI-CNN wins the 2nd-best performance in all the baseline approaches. Specifically, on MDAD dataset, its AUC and AUPRC values are 0.9332 and 0.9263, which are 2.5% and 2.2% lower than those of SCSMDA. On aBiofilm and DrugVirus datasets, its AUC values are 0.9467 and 0.8490, which is 1.9% and 4.4% lower than those of SCSMDA. Besides, on aBiofilm and DrugVirus datasets, Graph2MDA and NIMCGCN win the 2nd-best performance on AUPRC metric, and their values are 0.9485 and 0.8462, respectively. The results in Figure 4 fully demonstrate that SCSMDA is the most competitive approach in microbe–drug association prediction on these datasets.

Comparison with other baseline methods under different ratios

Different ratios between the number of positive samples and the number of negative samples may affect the performance of SCSMDA as well as the baseline approaches. Therefore, to evaluate their performance comprehensively, we conduct the evaluation experiments under three different ratios (# positive samples: # negative samples=1:1, 1:5 and 1:10, respectively) five times and obtain the mean and standard deviation values of the results. The corresponding results on AUC and AUPRC metrics are presented in Table 3.

For the result with the 1:1 ratio, SCSMDA wins the 1st rank on the three datasets. Specifically, the AUC and AUPRC values are 0.9573 and 0.9464 on MDAD dataset. Besides, the AUC and AUPRC values are 0.9658 and 0.9450 on aBiofilm dataset, whereas AUC and AUPRC values are 0.8834 and 0.8637 on DrugVirus, respectively. Meanwhile, DTI-CNN achieves the 2nd-best performance on these three datasets. Its AUC values are 0.9325, 0.9436 and 0.8581, and the AUPRC values are 0.9242, 0.9316 and 0.8396 on MDAD, aBiofilm and DrugVirus, respectively.

For the result with the 1:5 ratio, SCSMDA and DTI-CNN wins the 1st rank and 2nd rank on these three datasets. In particular, for the AUC metric, SCSMDA obtains the 0.9434, 0.9559 and 0.8757 scores, whereas DTI-CNN achieves the 0.9308, 0.9412 and 0.8466 scores, respectively. For the AUPCR metric, SCSMDA gets the 0.7607, 0.7971 and 0.5777 values, respectively, and DIT-CNN obtains 0.7545, 0.7891 and 0.5644 values, respectively.

For the result with the 1:10 ratio, SCSMDA achieves the highest scores on AUC metric, which are 0.9377, 0.9481 and 0.8729 on MADA, aBiofilm and DrugVirus datasets, respectively. Meanwhile, SCSMDA also achieves the best performance on AURPC metric for DrugVirus dataset with 0.4042. SCSMDA wins the 2nd-highest scores on AUPRC of MDAD and aBiofilm datasets and the values are 0.6920 and 0.6853. Besides, DTI-CNN wins the 1st rank on AUPRC metric for two datasets, and their AUPRC scores are 0.7071, and 0.6997 on MADA and aBiofilm datasets. DIT-CNN achieves the 2nd-best performance on AUC of MDAD, AUC of aBiofilm, AUC of DrugVirus and AUPRC of the DrugVirus, and their corresponding scores are 0.9356, 0.9332, 0.8469 and 0.3943, respectively. All the results are listed in Table 3, which comprehensively demonstrates that SCSMDA consistently has a better performance than other baseline approaches.

Model ablation study

SCSMDA learns the embedding of microbes and drugs with the structure-enhanced contrastive learning strategy, and selects the most informative samples with self-paced negative sampling

Table 3. The performance of SCSMDA for predicting microbe–drug associations under different ratios on MDAD, aBiofilm and DrugVirus datasets

		MDAD		aBiofilm		DrugVirus	
Ratios		**AUC**	**AUPRC**	**AUC**	**AUPRC**	**AUC**	**AUPRC**
1:1							
	GCN [44]	0.8631±0.0059	0.8668±0.0058	0.8878±0.0066	0.8873±0.0095	0.8202±0.0093	0.7985±0.0174
	GAT [51]	0.8755±0.0049	0.8772±0.0046	0.8995±0.0045	0.8922±0.0058	0.8033±0.0028	0.7908±0.0018
	DTIGAT [52]	0.9185±0.0023	0.9149±0.0066	0.9205±0.0024	0.9179±0.0041	0.8169±0.0102	0.8152±0.0105
	NIMCGCN [53]	0.8944±0.0087	0.9016±0.0068	0.9201±0.0066	0.9251±0.0051	0.8319±0.0065	0.8438±0.0468
	MMGCN [54]	0.8943±0.0022	0.9033±0.0051	0.9042±0.0032	0.9103±0.0056	0.7946±0.0110	0.7840±0.0139
	GCNMDA[13]	0.9299±0.0055	0.9192±0.0094	0.9407±0.0023	0.9291±0.0044	0.8330±0.0063	0.8047±0.0088
	DTI-CNN [55]	0.9325±0.0054	0.9242±0.0082	0.9436±0.0010	0.9316±0.0058	0.8581±0.0013	0.8396±0.0162
	SCSMDA (Ours)	**0.9573±0.0020**	**0.9464±0.0033**	**0.9658±0.0026**	**0.9450±0.0037**	**0.8834±0.0064**	**0.8637±0.0096**
1:5							
	GCN [44]	0.8830±0.0027	0.6829±0.0074	0.8808±0.0018	0.6715±0.0047	0.8291±0.0007	0.4845±0.0031
	GAT [51]	0.8717±0.0047	0.6325±0.0097	0.9021±0.0062	0.6867±0.0082	0.8169±0.0025	0.4725±0.0138
	DTIGAT [52]	0.9097±0.0003	0.7462±0.0056	0.9156±0.0042	0.7565±0.0060	0.8001±0.0022	0.4630±0.0058
	NIMCGCN [53]	0.8983±0.0039	0.7339±0.0051	0.9143 ± 0.0115	0.7626±0.0118	0.8424±0.0040	0.5280±0.0061
	MMGCN [54]	0.8964±0.0008	0.7295±0.0042	0.9072±0.0010	0.7584±0.0046	0.7791±0.0040	0.4764±0.0129
	GCNMDA [13]	0.9274±0.0006	0.7119±0.0082	0.9374±0.0043	0.7623±0.0441	0.8366±0.0054	0.4788±0.0156
	DTI-CNN [55]	0.9308±0.0015	0.7545±0.1011	0.9412±0.0006	0.7891±0.0014	0.8466±0.0006	0.5644±0.0045
	SCSMDA (Ours)	**0.9434±0.0048**	**0.7607±0.0193**	**0.9559±0.0026**	**0.7971±0.0041**	**0.8757±0.0003**	**0.5777±0.0046**
1:10							
	GCN [44]	0.8921±0.0065	0.5821±0.0170	0.8974±0.0018	0.5879±0.0035	0.8231±0.0018	0.3255±0.0065
	GAT [51]	0.8696±0.0017	0.5324±0.0073	0.8999±0.0015	0.5828±0.0103	0.8089±0.0023	0.3208±0.0094
	DTIGAT [52]	0.9085±0.0064	0.6483±0.0264	0.9156±0.0010	0.6419±0.0091	0.7957 ± 0.0012	0.3068±0.0022
	NIMCGCN [53]	0.9009±0.0008	0.6256±0.0108	0.9119±0.0022	0.6579±0.0030	0.8414±0.0074	0.3503±0.0076
	MMGCN [54]	0.8941±0.0011	0.6244±0.0031	0.9044±0.0005	0.6463±0.0028	0.7765±0.0048	0.3596±0.0086
	GCNMDA[13]	0.9310±0.0028	0.5939±0.0234	0.9415±0.0010	0.6201±0.0033	0.8304±0.0055	0.3139±0.0139
	DTI-CNN [55]	0.9356±0.0011	**0.7071±0.0010**	0.9332±0.0017	**0.6997±0.0081**	0.8649±0.0020	0.3943±0.0080
	SCSMDA (ours)	**0.9377±0.0015**	0.6921±0.0069	**0.9481±0.0009**	0.6853±0.0049	**0.8729±0.0017**	**0.4042±0.0016**

Note: The best results are marked in bold and the 2nd-best ones are marked as underlined.

strategy. Here we conduct the model ablation study to investigate the effect of each component on SCSMDA model. Here we mainly select three components which are the similarity-network-based embedding learning component (SN), the meta-path-induced network embedding learning component (MP) and the self-paced negative sampling strategy component (SP). The ablation study is performed as SCSMDA without SN component (SCS w/o SN), SCSMDA without MP component (SCS w/o MP), SCSMDA without SP component (SCS w/o SP) and SCSMDA with all these components. The corresponding results are represented in Figure 5.

Results on all these three datasets show that SN, MP and SP are all essential components for SCSMDA. Specifically, on MDAD dataset, SCSMDA wins the best performance on the five evaluation metrics. On MDAD dataset, the scores on ACC, AUC, AUPRC, MCC and F1 metric are 0.8791, 0.9573, 0.9464, 0.7261 and 0.8528, respectively. For aBiofilm dataset, the scores of ACC, AUC, AUPRC, MCC and F1 metrics are 0.8919, 0.9658, 0.9450, 0.7393, and 0.8592, respectively. On DrugVirus dataset, the values on ACC, AUC, AUPRC, MCC and F1 metrics are 0.8133, 0.8834, 0.8637, 0.6141 and 0.7981, respectively.

For other prediction models, SCSMDA w/o SP achieves the 2nd-best performance overall, whereas the performance of SCSMDA w/o SN model is the worst in all the models. The corresponding results for other modes are displayed in Figure 5 and we don't repeat them anymore. Overall, the embedding of nodes learning from the similarity-network-based plays a major role in the performance of SCSMDA. Meanwhile, the structure-enhanced learning component plays an essential role in improving the performance of SCSMDA. The structure-enhanced contrasting learning strategy is effective in improving the performance of SCSMDA.

The statistical significance report on AUC values

The statistical significance is an effective manner for verifying the credibility and stability of the results of SCSMDA. Therefore, we employ the one-way ANOVA model [56, 57] to investigate the statistical significance of the results of all the MDA prediction approaches. Specifically, all these MDA prediction approaches are performed on the 5-CV experiments and obtain their corresponding AUC values (Table 4). The analysis results are demonstrated in Figure 6.

The results show that the *P*-values between SCSMDA and other baseline approaches (GCNMDA, GCN, GAT, DTI-GAT, NIMCGCN, MMGCN, DTI-CNN and Graph2MDA) are 9.9e-7, 6.2e-12, 6.5e-07, 3.4e-04, 9.9e-9, 7.3e-12, 2.2e-7 and 1.1e-4 on MDAD datasets, which all show SCSMDA has statistical significance values according to one-way ANOVA analysis. Besides, we also display the *P*-values between baseline approaches. The statistical significance analysis results on aBiofilm and DrugVirus are all displayed in Figure 6B and C and we don't repeat them anymore.

Embedding size analysis on SCSMDA

SCSMDA learns the embeddings of microbes and drugs with the structure-enhanced contrastive learning strategy. Since the embedding size of microbes and drugs plays an important role in SCSMDA, we conduct this experiment and evaluate its impact on SCSMDA with five metrics which are ACC, AUC, AUPRC, MCC and F1. Here, we set the embeddings size of microbes and drugs as 32,

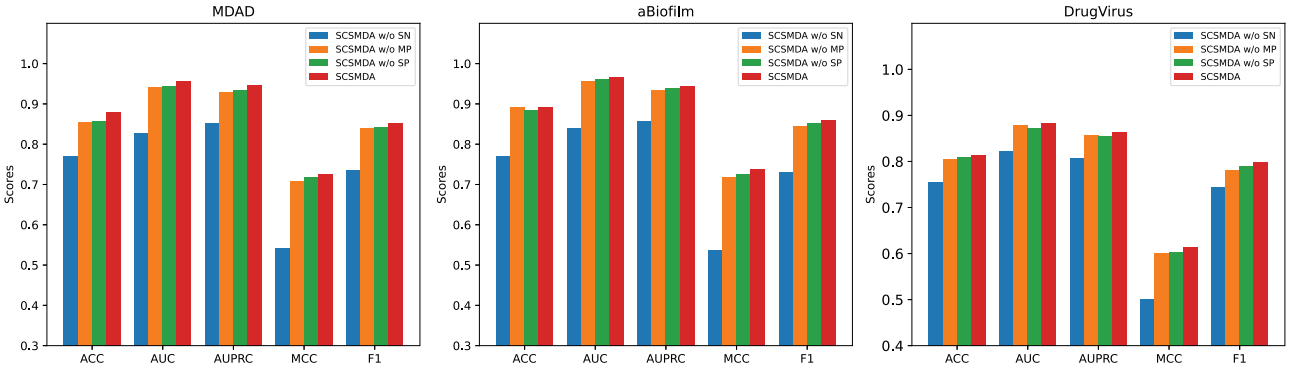

Figure 5. The ablation study for SCSMDA. SCSMDA w/o SN, SCSMDA w/o MP and SCSMDA w/o SP indicate that SCSMDA doesn't contain similarity-network-based embedding learning component, meta-path-induced network embedding learning component and the self-paced negative sampling strategy component, respectively.

Table 4. AUC values of baseline approaches under the 5-CV experiment on each dataset

Dataset	Iteration	GCN	GAT	DTIGAT	NIMCGCN	MMGCN	GCNMDA	DTI-CNN	Graph2MDA	SCSMDA (ours)
MDAD	1	0.8685	0.8873	0.8692	0.8892	0.8934	0.9326	0.9326	0.9077	0.9562
	2	0.8715	0.8899	0.9134	0.9001	0.8938	0.9261	0.9361	0.9022	0.9583
	3	0.8702	0.8856	0.9136	0.9018	0.8940	0.9297	0.9358	0.8617	0.9603
	4	0.8729	0.8695	0.9145	0.899	0.8937	0.9328	0.9319	0.9089	0.9563
	5	0.8738	0.8616	0.9139	0.8933	0.8941	0.9280	0.9303	0.8756	0.9617
aBiofilm	1	0.8987	0.8962	0.9192	0.9009	0.9083	0.9382	0.9443	0.9164	0.9667
	2	0.8997	0.8758	0.9196	0.9117	0.9077	0.9398	0.9454	0.9212	0.9614
	3	0.9009	0.8898	0.9207	0.9147	0.9081	0.9424	0.9448	0.9125	0.9661
	4	0.8999	0.9038	0.9206	0.9193	0.9084	0.9412	0.9427	0.8894	0.9664
	5	0.9031	0.9048	0.9198	0.8964	0.9082	0.9422	0.9406	0.9272	0.9669
DrugVirus	1	0.8349	0.8036	0.8184	0.8427	0.7931	0.8349	0.8612	0.7725	0.8934
	2	0.8353	0.7956	0.8203	0.8415	0.7937	0.7901	0.8611	0.7981	0.8841
	3	0.8356	0.7959	0.8190	0.8440	0.7962	0.8413	0.8566	0.7802	0.8845
	4	0.8349	0.7876	0.8230	0.8372	0.8237	0.8264	0.8574	0.7899	0.8888
	5	0.8349	0.7902	0.8164	0.8346	0.8215	0.8171	0.8611	0.7991	0.8865

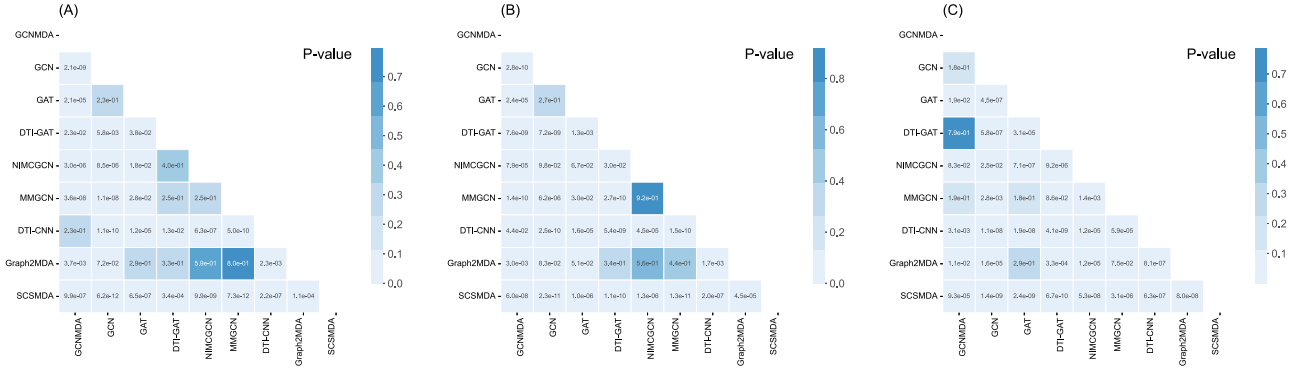

Figure 6. The statistical significance report with one-way ANOVA model. (A) *P*-values on MDAD dataset, (B) *P*-values on aBiofilm dataset,(C) *P*-values on DrugVirus dataset.

62, 128, 256 and 512, respectively, and the corresponding results are shown in Table 5.

Specifically, on MDAD dataset, the ACC, AUC, AUPRC and F1 values are 0.8719, 0.9573, 0.9464 and 0.8528, which are the highest scores when the embedding size is 128. The highest score on MCC is 0.7365 when the embedding size is 64. For aBiofilm dataset, the highest scores for ACC, AUC, MCC and F1 are 0.8919, 0.9658, 0.7393 and 0.8592 when the embedding size is 128 and the highest value for AUPRC is 0.9458 when the embedding size is 64. For DrugVirus dataset, SCSMDA performs best on ACC, AUC, AUPRC, MCC and F1 when the embedding size is 64, 64, 128, 256 and 128, respectively. From the results, we can find that the embedding size affects the performance of SCSMDA model. SCSMDA achieves the highest scores when the embedding size is 128 overall. As a result, we adopt the embedding size as 128 for SCSMDA.

Parameter sensitivity analysis

For SCSMDA model, some crucial parameters affect its performance. Here we mainly focus on five parameters: the number of positive pairs, the number of GCN layers, the number of MLP layers, the number of bins and the learning rate. The corresponding

Table 5. The performance of SCSMDA under different embedding sizes on MDAD, aBiofilm and DrugVirus datasets.

Dataset	Embedding size	ACC	AUC	AUPRC	MCC	F1
MDAD	16	0.8582± 0.0052	0.9478±0.0031	0.9243±0.0045	0.7329±0.0121	0.8409±0.0027
	32	0.8659±0.0036	0.9506±0.0019	0.9364±0.0044	0.7352±0.0062	0.8485±0.0027
	64	0.8701±0.0045	0.9548±0.0022	0.9409±0.0027	**0.7365±0.0061**	0.8504±0.0036
	128	**0.8791±0.0054**	**0.9573±0.0020**	**0.9464±0.0033**	0.7261±0.0025	**0.8528±0.0008**
	256	0.8651±0.0030	0.9511±0.0043	0.9389±0.0068	0.7008±0.0249	0.8477±0.0181
	512	0.8304±0.0131	0.9446±0.0035	0.9330±0.0044	0.7093±0.0156	0.8491±0.0095
aBiofilm	16	0.8824±0.0013	0.9538±0.0028	0.9491±0.0082	0.7316±0.0013	0.8627±0.0081
	32	0.8907±0.0077	0.9633±0.0011	0.9430±0.0029	0.7384±0.0161	0.8590±0.0112
	64	0.8915±0.0070	0.9644±0.0041	**0.9485±0.0049**	0.7367±0.0077	0.8576±0.0037
	128	**0.8919±0.0017**	**0.9658±0.0026**	0.9450±0.0037	**0.7393±0.0041**	**0.8592±0.0031**
	256	0.8864±0.0029	0.9632±0.0003	0.9426±0.0006	0.7317±0.0060	0.8542±0.0035
	512	0.8762±0.0072	0.9560±0.0085	0.9388±0.0071	0.7249±0.0004	0.8371±0.0007
DrugVirus	16	0.8071±0.0100	0.8748±0.0088	0.8469±0.0121	0.6002±0.0172	0.7845±0.0059
	32	0.8165±0.0132	0.8843±0.0007	0.8575±0.0109	0.6027±0.0117	0.7899±0.0048
	64	**0.8196±0.0080**	**0.8861±0.0110**	0.8572±0.0173	0.6109±0.0272	0.7955±0.0148
	128	0.8133±0.0082	0.8834±0.0064	**0.8637±0.0096**	0.6141±0.0063	**0.7981±0.0016**
	256	0.8096±0.0032	0.8769±0.0028	0.8611±0.0069	**0.6218±0.0092**	0.7979±0.0076
	512	0.8031±0.0024	0.8713±0.0026	0.8624±0.0014	0.5974±0.0212	0.7881±0.0121

Note: The best results are marked in bold.

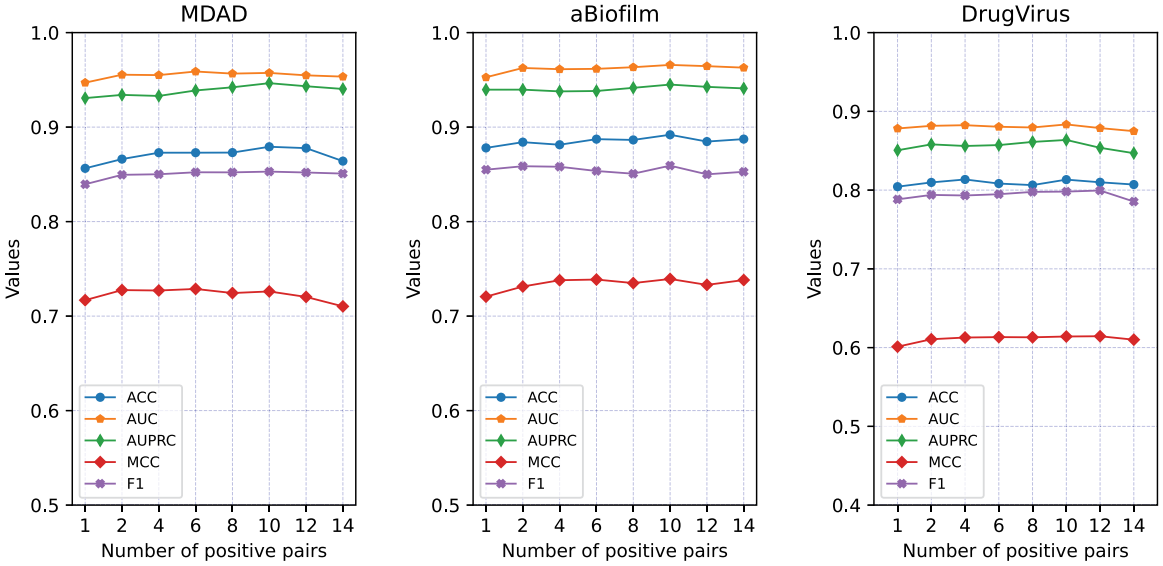

Figure 7. The performance of SCSMDA under different numbers of positive pairs on MDAD, aBiofilm and DrugVirus datasets.

experiments are performed and the results are all evaluated with ACC, AUC, AUPRC, MCC and F1.

The 1st parameter is the number of positive pairs for structure-enhanced contrastive learning strategy. We vary the number of positive pairs from {1,2,4,6,8,10,12,14} and conduct the experiments on all three datasets. The results are presented in Figure 7. Specifically, on the MDAD dataset, the values of ACC, AUC, AUPRC, MCC and F1 first increase gradually and then slightly decreases with positive sample number ranging from {1,2,4,6,8,10,12,14}. When the threshold is 10, the scores are highest and the values are 0.8791, 0.9573, 0.9464, 0.7261 and 0.8528 on ACC, AUC, AUPRC, MCC and F1, respectively. For aBiofilm and DrugVirus datasets, their results are similar to those on MDAD dataset and we don't repeat them anymore. It should be noted that the evaluation scores are almost the lowest when the number of positive pairs is 1. This could further confirm that our novel positive-pair selection strategy is helpful in improving the performance of SCSMDA. As a result, we set the number of positive pairs as 10.

The 2nd parameter is the number of the MLP layer. MLP is employed as the classifier to predict MDAs, which directly affects the performance of the SCSMDA. It is very critical to choose a proper layer number for MLP. The corresponding results (Figure 8) fully indicate that SCSMDA achieves the best performance when the number of the MLP layer is 1. Previous studies also find that too many MLP layers may lead to over-smoothing [58, 59], which seriously affects the performance of the prediction model. SCSMDA achieves its best results when the number of MLP layers is 1, which is consistent with the previous study. The 3rd parameter is the number of the GCN layer. GCN is employed to learn the embeddings of microbes and drugs, which is decisive to the prediction accuracy of SCSMDA. The results under different GCN lay numbers are presented in Figure 8 The best performance is achieved when the number of GCN layers is 1.

The last two parameters are the learning rate and the number of bins. The learning rate is a hyperparameter that controls how much to change one model in response to the estimated error

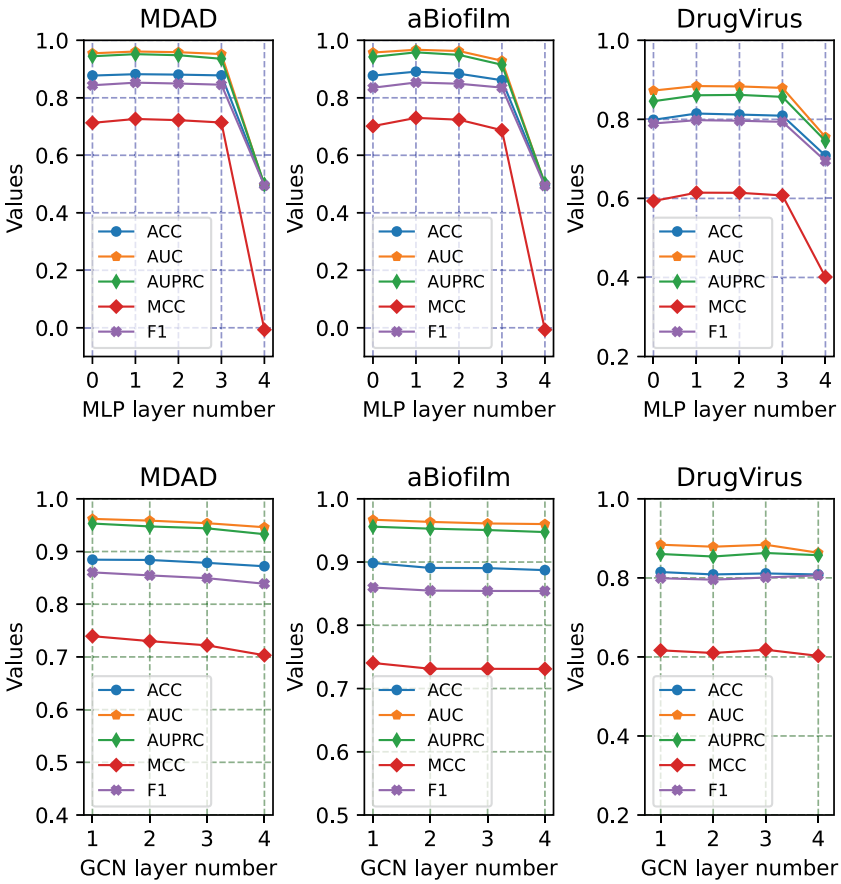

Figure 8. The performance of SCSMDA under different numbers of MLP layers and GCN layers on MDAD, aBiofilm and DrugVirus datasets.

[60]. Choosing a proper learning rate is challenging, since a small value may result in a long training process, while a too-large value may result in learning an unstable training process. As a result, SCSMDA searches on learning rate from {1e-2, 1e-3, 5e-3, 1e-4, 5e-4, 1e-5} and we evaluate the performance of SCSMDA under these different learning rates. The results are shown in Figure 9. We observe that performance of SCSMDA first increases and then slightly decreases with the weights from 1e-1 to 1e-5. SCSMDA achieves the best results when the learning rate is 5e-4. Lastly, for the number of bins which is the hyperparameter in self-paced negative sampling strategy process, SCSMDA chooses the values from {2, 4, 6, 8, 10,12} and the corresponding results are presented in Figure 9. SCSMDA obtains the best scores when the number of bins equals 10.

Visualization and interpretation for the embeddings of microbe–drug pairs learned by SCSMDA

To further demonstrate the outstanding ability of SCSMDA in learning the embedding of nodes, we conduct the visualization experiment on aBiofilm dataset. Specifically, with the learned embeddings of microbes and drugs, novel embeddings for microbe–drug pairs are generated based on the Hadamard products. If one microbe and one drug have an association relationship, this microbe–drug pair will be labeled with a positive pair. Otherwise, it will be labeled with a negative pair. All the embeddings of microbe–drug pairs are plotted into a two-dimensional space using t-SNE tool [61]. The visualization results are displayed in Figure 10.

It can be seen that the positive pairs and the negative pairs are gradually distinguished with the increase of the epochs. The embeddings of positive pairs and the negative pairs are in chaos when the epoch number is 1. The embedding distribution is gradually clear with the epochs increase. Finally, the positive pairs (red points) and the negative pairs (blue points) are almost separated when the epochs equal 100. Meanwhile, it should be noted that some red and green dots are still mixed in some areas, indicating that the decision boundary is very difficult in microbe–drug association prediction task. This observation further confirms that the learned embeddings of microbe–drug pairs are discriminative and interpretable, which improves the accuracy of SCSMDA in predicting MDAs.

Running time of SCSMDA and baseline approaches

To fully evaluate the execution efficiency of SCSMDA as well as the comparison approaches, we conduct the 5-CV experiment on the three datasets for each prediction model and compare their corresponding running time. The 5-CV experiments were conducted five times independently and their corresponding results are all displayed in Table 6.

The results indicate that method DIT-CNN requires the shortest running time, whereas Graph2MDA needs the longest running time. The average running time on MDAD, aBiofilm and DrugVirus datasets for DIT-CNN is 10, 10 and 4s. The average running time on MDAD, aBiofilm and DrugVirus datasets for Graph2MDA is 788, 1261 and 52s. For our proposed model SCSMDA, its average running time on MDAD, aBiofilm and DrugVirus is 342, 450 and

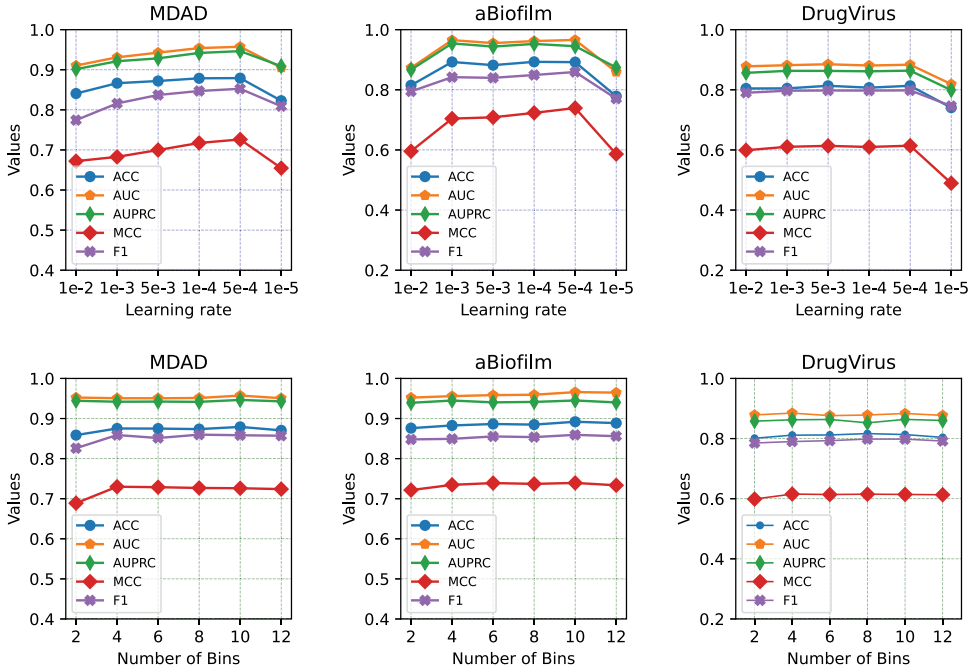

Figure 9. The performance of SCSMDA under different thresholds for learning rate and number of Bins on MDAD, aBiofilm and DrugVirus datasets.

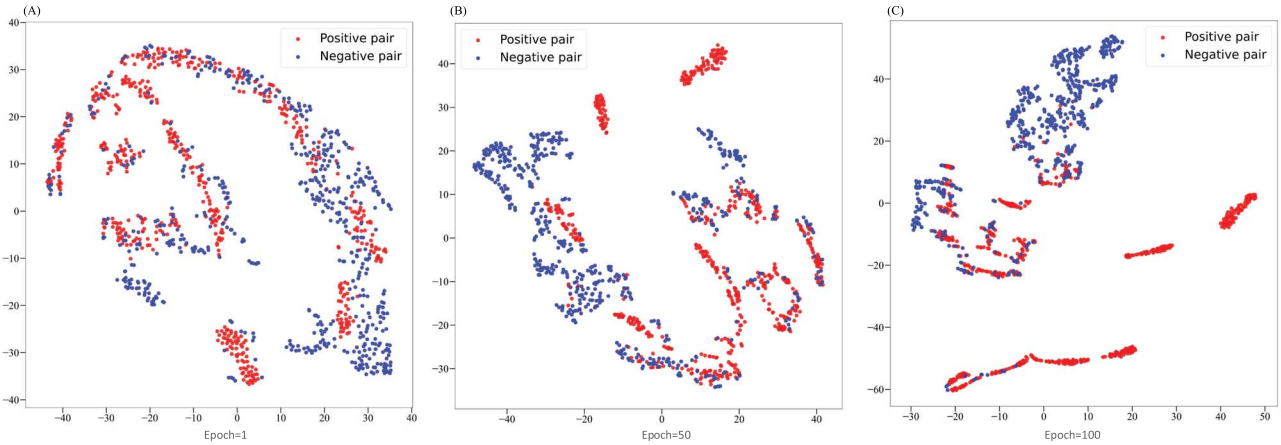

Figure 10. Visualization of the learned microbe–drug embeddings by SCSMDA on aBiofilm under different epochs.

134s, respectively. The results illustrate that our proposed method could complete training and prediction tasks within an acceptable time.

Case study

To comprehensively verify the ability of SCSMDA in finding novel MDAs, we perform case studies on two popular antimicrobial drugs ciprofloxacin and moxifloxacin, which is the same as the previous research [15]. Specifically, for each target drug, all the known microbe–drug associations will be set to unknown, and then all the candidate microbes will be sorted in a descending manner according to their scores predicted by SCSMDA. Lastly, we screen out the top-20 ranked microbes and verify them by published literature. The case study results for ciprofloxacin and moxifloxacin are displayed in Tables 7 and 8.

Drug ciprofloxacin belongs to a class of drugs called quinolone antibiotics. It usually is used to treat a variety of bacterial infections such as urinary tract infections and pneumonia [62]. Previous studies have indicated that ciprofloxacin has a close relationship with many human microbes. For example, it is reported that *Candida albicans* and *Staphylococcus aureus* together could result in biofilm formation and increase antimicrobial resistance. Daniel [63] fully accessed the susceptibility between ciprofloxacin and Salmonella and found that ciprofloxacin susceptibility was highly dependent on serotype. Besides, Mercedes [64] discovered that the activity of ciprofloxacin against *Bacillus subtilis* species depends on the drug's interaction with its target enzymes. The results for other predicted microbes are displayed in Table 7 and 16 out of top 20 predicted candidate microbes related to ciprofloxacin can be confirmed by literature.

Drug moxifloxacin is also a common antibiotic, which is always employed to treat bacterial infections including pneumonia, conjunctivitis, endocarditis, tuberculosis and sinusitis [65, 66]. Moxifloxacin could inhibit the reproduction growth rate and life cycle of broad-spectrum bacteria. For example, *Escherichia coli* is a bacteria that normally lives in the intestines of both healthy people and animals. Axel [67] suggested that moxifloxacin had a potential impact on bactericidal activities of *Escherichia coli*. *Staphylococcus aureus* is a Gram-positive spherically shaped bacterium, a member of the Bacillota. Dilek [68] stated that moxifloxacin had enhanced

Table 6. Running time (seconds) of SCSMDA and other baseline approaches on MDAD, aBiofilm and DrugVirus datasets.

Datasets	Rounds	GCN	GAT	DTIGAT	NIMCGCN	MMGCN	GCNMDA	DTI-CNN	Graph2MDA	SCSMDA(ours)
MDAD	1	98	293	296	114	116	303	10	786	342
	2	109	296	295	121	114	302	10	788	341
	3	106	299	289	118	115	302	9	790	346
	4	100	296	295	121	117	303	10	788	340
	5	109	297	295	119	116	311	9	787	340
	AVE	104	296	294	118	116	304	10	788	342
aBiofilm	1	127	393	379	161	170	399	11	1255	417
	2	143	387	381	162	147	394	9	1266	589
	3	142	386	385	164	147	394	10	1256	418
	4	144	386	383	187	148	395	11	1261	417
	5	145	387	382	163	147	395	10	1269	411
	AVE	140	388	382	167	152	395	10	1261	450
DrugVirus	1	16	69	69	19	17	28	4	50	132
	2	15	71	66	19	16	28	4	53	136
	3	14	71	68	17	17	28	4	53	135
	4	15	66	67	18	17	28	4	53	136
	5	15	70	70	17	16	28	4	52	130
	AVE	15	70	68	18	17	28	4	52	134

Note: AVE denotes the average running time of the five 5-CV experiment for each model.

Table 7. The top-20 predicted Ciprofloxacin-associated microbes by SCSMDA

Microbe name	Rank	Evidence	Microbe name	Rank	Evidence
Candida albicans	1	PMID:31471074	*Listeria monocytogenes*	11	PMID:28355096
Streptococcus mutans	2	PMID:30468214	*Bacillus cereus*	12	PMID:8448312
Salmonella enterica	3	PMID:26933017	*Burkholderia pseudomallei*	13	PMID:24502667
Staphylococcus epidermidis	4	PMID:28481197	*Streptococcus epidermidis*	14	Unconfirmed
Burkholderia cenocepacia	5	PMID:27799222	*Campylobacter jejuni*	15	PMID:11920303
Bacillus subtilis	6	PMID:15194135	*Agrobacterium tumefaciens*	16	Unconfirmed
Serratia marcescens	7	PMID:23751969	*Vibrio vulnificus*	17	PMID:24978586
Acinetobacter baumannii	8	PMID:25147676	*Staphylococcus epidermidis*	18	PMID:10632381
Streptococcus sanguis	9	PMID:11347679	*Candida tropicalis*	19	Unconfirmed
Vibrio harveyi	10	PMID:27247095	*Actinomyces oris*	20	Unconfirmed

Table 8. The top-20 predicted Moxifloxacin-associated microbes by SCSMDA

Microbe name	Rank	Evidence	Microbe name	Rank	Evidence
Escherichia coli	1	PMID:31542319	*Burkholderia cenocepacia*	11	PMID:28355096
Streptococcus mutans	2	PMID:29160117	*Serratia marcescens*	12	Unconfirmed
Staphylococcus aureus	3	PMID:12654680	*Burkholderia pseudomallei*	13	PMID:24502667
Pseudomonas aeruginosa	4	PMID:31691651	*Streptococcus epidermidis*	14	Unconfirmed
Staphylococcus epidermidis	5	PMID:11249827	*Acinetobacter baumannii*	15	PMID:12951327
Vibrio harveyi	6	Unconfirmed	*Salmonella enterica*	16	PMID:22151215
Staphylococcus epidermidis	7	PMID:31516359	*Vibrio cholerae*	17	Unconfirmed
Enterococcus faecalis	8	PMID:31763048	*Vibrio vulnificus*	18	PMID:10632381
Listeria monocytogenes	9	PMID:28739228	*Klebsiella pneumoniae*	19	PMID:27257956
Proteus mirabilis	10	PMID:15077996	*Actinomyces oris*	20	Unconfirmed

potency against S. aureus. Besides, some studies confirmed that bactericidal activity of moxifloxacin is against S. aureus strains *in vitro* [69]. We display the top-20 predicted candidate microbes in Table 8 and 15 of them can be verified by previous publications. Case studies on these two drugs further indicate that SCSMDA has a good performance in identifying novel MDAs.

Besides, SCSMDA conducts the case study for each microbe and drug on the three public datasets. The correspondence results are available in the GitHub and we don't repeat them anymore.

Conclusion

Recent studies have comprehensively shown that microbes residing within and upon human bodies play critical roles in human health. Accurately identifying the microbe–drug associations is a crucial step in precision medicine. Here we propose a novel approach named SCSMDA to predict microbe–drug associations which achieves the best performance among all the baseline approaches. SCSMDA employs the meta-path-induced networks of microbes and drugs to enhance their feature representations

learned from the similarity networks with the contrastive learning strategy, which could obtain their deep-level representations. Besides, SCSMDA utilizes the self-paced negative sampling strategy to select the most informative negative samples for training the MLP classifier more efficiently.

To comprehensively evaluate the performance of SCSMDA as well as the baseline methods, we conduct the 5-CV experiment on three public datasets. Experimental results show that the proposed method wins the highest scores on the AUC and AUPRC evaluation metrics. We also conduct the comparison experiments under different ratios (# positive sample: # negative samples=1:1, 1:5 and 1:10). SCSMDA achieves the best performance on these datasets. Besides, the model ablation experiment is adopted to further verify the effectiveness of the structure-enhanced contrastive learning strategy and self-paced negative sampling strategy. Meanwhile, parameter sensitivity experiments are employed to tune the best parameters for SCSMDA. In the end, the results of case studies on two common drugs could be supported by published literature, which further confirms the advantages of SCSMDA in discovering novel MDAs.

Next, we can do some work from the following two aspects. Firstly, some other biological entities such as genes and proteins could be employed to establish a more comprehensive knowledge graph related to microbes and drugs. We can learn the embedding of microbes and drugs with the help of knowledge graphs aiming to improve the prediction accuracy of the MDA prediction model. Secondly, since association relationship predictions between biological entities are one of the foundation tasks in computational biology, we can apply SCSMDA to other link prediction problems such as drug-drug interaction and miRNA–disease association prediction.

Key Points

- SCSMDA constructs the meta-path induced networks for microbes and drugs by utilizing their different meta-paths with semantic meanings.
- SCSMDA employs the structure-enhanced contrastive learning strategy to obtain the effective representations of microbes and drugs.
- SCSMDA adopts the self-paced negative sampling strategy to select the most informative negative samples for training the MLP classifier.
- Results on these three datasets comprehensively indicate that SCSMDA outperforms seven other baseline methods in microbe–drug association prediction task.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Funding

National Science Foundation of China (No. 61801432, 62031003).

Author contributions statement

Z.T. conceived the experiment and the whole manuscript. Y.Y. developed the codes and algorithm. Z.T., H.F. and Y.Y. set up the general idea of this study. W.X. and M.G. revised the manuscript. All authors have read and approved the manuscript.

Availability and implementation

The source code and databases are available at https://github.com/Yue-Yuu/SCSMDA-master.

References

1. Human Microbiome Project Consortium, *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**(7402): 207–14.
2. Ventura M, Oflaherty S, Claesson MJ, *et al.* Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat Rev Microbiol* 2009; **7**(1): 61–71.
3. Kau AL, Ahern PP, Griffin NW, *et al.* Human nutrition, the gut microbiome and the immune system. *Nature* 2011; **474**(7351): 327–36.
4. Sommer F, Bäckhed F. The gut microbiota-masters of host development and physiology. *Nat Rev Microbiol* 2013; **11**(4): 227–38.
5. Zhang H, John K, Baise D, *et al.* Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci*, **106**(7): 2365–70, 2009.
6. Wen L, Ley RE, Volchkov PY, *et al.* Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature* 2008; **455**(7216): 1109–13.
7. Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer* 2013; **13**(11): 800–12.
8. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, *et al.* Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 2019; **570**(7762): 462–7.
9. Guthrie L, Gupta S, Daily J, *et al.* Human microbiome signatures of differential colorectal cancer drug metabolism. *NPJ Biofilms Bicrobiomes* 2017; **3**(1): 1–8.
10. Kashyap PC, Chia N, Nelson H, *et al.* Microbiome at the frontier of personalized medicine. In *Mayo Clinic Proceedings*, Vol. **92**. Elsevier, 2017, 1855–64.
11. Long Y, Min W, Liu Y, *et al.* Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics* 2022; **38**(8): 2254–62.
12. Zhu L, Duan G, Yan C, *et al.* Prediction of microbe-drug associations based on chemical structures and the katz measure. *Curr Bioinform* 2021; **16**(6): 807–19.
13. Long Y, Min W, Kwoh CK, *et al.* Predicting human microbe–drug associations via graph convolutional network with conditional random field. *Bioinformatics* 2020; **36**(19): 4918–27.
14. Long Y, Luo J. Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. *IEEE J Biomed Health Inform* 2020; **25**(1): 266–75.
15. Long Y, Min W, Liu Y, *et al.* Ensembling graph attention networks for human microbe–drug association prediction. *Bioinformatics* 2020; **36**(Supplement_2): i779–86.
16. Deng L, Huang Y, Liu X, *et al.* Graph2mda: a multi-modal variational graph embedding model for predicting microbe–drug associations. *Bioinformatics* 2022; **38**(4): 1118–25.
17. Yang H, Ding Y, Tang J, *et al.* Inferring human microbe–drug associations via multiple kernel fusion on graph neural network. *Knowl Based Syst* 2022; **238**:107888.
18. Liu X, Zhang F, Hou Z, *et al.* Self-supervised learning: generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2023; **35**(1):857–76.

19. Hassani K, and Khasahmadi AH,. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–26. PMLR, 2020.
20. Peng Z, Huang W, Luo M, *et al.* Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*, WWW '20, page259–270, 2020. New York, NY, USA: Association for Computing Machinery.
21. Li Y, Qiao G, Gao X, *et al.* Supervised graph co-contrastive learning for drug-target interaction prediction. *Bioinformatics* 2022; **38**(10): 2847–54 03.
22. Wang R, Jin J, Zou Q, *et al.* Predicting protein-peptide binding residues via interpretable deep learning. *Bioinformatics* 2022; **1**:10.
23. Liu X, Song C, Huang F, *et al.* GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief Bioinform* 2021; **23**(1) 11:bbab457.
24. Wang Y, Min Y, Chen X, *et al.* Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the Web Conference 2021*, WWW '21, page2921–2933, 2021. New York, NY, USA: Association for Computing Machinery.
25. Wang X, Liu N, Han H, *et al.* Self-supervised heterogeneous graph neural network with co-contrastive learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1726–36, New York, NY, USA, 2021, Association for Computing Machinery.
26. Sheng Wan, Shirui Pan, Jian Yang, and Chen Gong. Contrastive and generative graph convolutional networks for graph-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume **35**, pages 10049–57, a Virtual Conference, 2021.
27. Lirong W, Lin H, Tan C, *et al.* Self-supervised learning on graphs: contrastive, generative,or predictive. *IEEE Transactions on Knowledge and Data Engineering* 2021;1–1.
28. Li F, Dong S, Leier A, *et al.* Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Brief Bioinform* 2022; **23**(1): bbab461.
29. Yang P, Li X-L, Mei J-P, *et al.* Positive-unlabeled learning for disease gene identification. *Bioinformatics* 2012; **28**(20): 2640–7.
30. Lou Z, Cheng Z, Li H, *et al.* Predicting miRNA-disease associations via learning multimodal networks and fusing mixed neighborhood information. *Brief Bioinform* 2022; **23**(5) 05:bbac159.
31. Jiang L, Sun J, Wang Y, *et al.* Identifying drug–target interactions via heterogeneous graph attention networks combined with cross-modal similarities. *Brief Bioinform* 2022; **23**(2): bbac016.
32. Kaiyang Q, Wei L, Zou Q. A review of dna-binding proteins prediction methods. *Curr Bioinform* 2019; **14**(3): 246–54.
33. Zhao T, Yang H, Valsdottir LR, *et al.* Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 2021; **22**(2): 2141–50.
34. Ding Y, Tang J, Guo F, *et al.* Identification of drug–target interactions via multiple kernel-based triple collaborative matrix factorization. *Brief Bioinform* 2022; **23**(2).
35. López V, Fernández A, García S, *et al.* An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inform Sci* 2013; **250**:113–41.
36. Zeng X, Zhong Y, Lin W, *et al.* Predicting disease-associated circular rnas using deep forests combined with positive-unlabeled learning methods. *Brief Bioinform* 2020; **21**(4): 1425–36.
37. Dai Q, Wang Z, Liu Z, *et al.* Predicting mirna-disease associations using an ensemble learning framework with resampling method. *Brief Bioinform* 2022; **23**(1): bbab543.
38. Wei H, Xu Y, Liu B. Ipidi-pul: identifying piwi-interacting rna-disease associations based on positive unlabeled learning. *Brief Bioinform* 2021; **22**(3): bbaa058.
39. Sun Y-Z, Zhang D-H, Cai S-B, *et al.* Mdad: a special resource for microbe-drug associations. *Frontiers in cellular andinfection microbiology*, 2018;**8**:424.
40. Rajput A, Thakur A, Sharma S, *et al.* Abiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Res* 2018; **46**(D1): D894–900.
41. Andersen PI, Ianevski A, Lysvand H, *et al.* Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int J Infect Dis* 2020; **93**:268–76.
42. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS Comput Biol* 2017; **13**(2): e1005366.
43. Hattori M, Tanaka N, Kanehisa M, *et al.* Simcomp/subcomp: chemical structure search servers for network analyses. *Nucleic Acids Res* 2010; **38**(suppl_2): W652–6.
44. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907* 2016.
45. Sixiao Zhang, Hongxu Chen, Xiangguo Sun, Yicong Li, and Guandong Xu.Unsupervised graph poisoning attack via contrastive loss back-propagation. In *Proceedings of the ACM Web Conference 2022*, pages 1322–30, 2022, New York, NY, USA: Association for Computing Machinery.
46. Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. Self-paced ensemble for highly imbalanced massive data classification. 2020 *IEEE 36th International Conference on DataEngineering (ICDE)*, Dallas, TX, USA, 2020, pp. 841–852, doi:https://doi.org/10.1109/ICDE48307.2020.00078.
47. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 249–256,Chia Laguna Resort, Sardinia, Italy. JMLR Workshop and Conference Proceedings*, 2010.
48. Kingma DP, Adam JB. A method for stochastic optimization. *arXiv preprint arXiv:14126980* 2014.
49. Jiajie Peng, Yuxian Wang, Jiaojiao Guan, Jingyi Li, Ruijiang Han, Jianye Hao, Zhongyu Wei, and Xuequn Shang. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Brief Bioinform*, **22**(5): bbaa430, 2021.
50. Tian Z, Peng X, Fang H, *et al.* MHADTI: predicting drug–target interactions via multiview heterogeneous information network embedding with hierarchical attention mechanisms. *Brief Bioinform* 2022; **23**(6): bbac434.
51. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. *arXiv preprint arXiv:171010903* 2017.
52. Wang H, Zhou G, Liu S, *et al.* Drug-target interaction prediction with graph attention networks. *arXiv preprint arXiv:210706099* 2021.
53. Li J, Zhang S, Liu T, *et al.* Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. *Bioinformatics* 2020; **36**(8): 2538–46.
54. Tang X, Luo J, Shen C, *et al.* Multi-view multichannel attention graph convolutional network for mirna–disease association prediction. *Brief Bioinform* 2021; **22**(6): bbab174.
55. Peng J, Li J, Shang X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics* 2020; **21**(13): 1–13.

56. Vijayvargiya A. One-way analysis of variance. *Journal of Validation Technology* 2009; **15**(1): 62.
57. Quirk TJ. One-way analysis of variance (anova). *Excel 2007 for educational and psychological Statistics*. New York, NY: Springer, 2012; 163–79.
58. Sun F. Over-smoothing effect of graph convolutional networks. *arXiv preprint arXiv:220112830* 2022.
59. Yang C, Wang R, Yao S, *et al.* Revisiting over-smoothing in deep gcns. *arXiv preprint arXiv:200313663* 2020.
60. Brownlee J. Understand the impact of learning rate on neural network performance. *Mach Learn Mastery* 2019;**20**: 1–27.
61. Van der Maaten L, Hinton G. Visualizing data using t-sne. *Journalof Machine Learning Research* 2008;**9**(11): 2579–2605.
62. Thai T, Salisbury BH, Zito PM. *Ciprofloxacin*StatPearls [Internet]. StatPearls Publishing, 2021.
63. Eibach D, Al-Emran HM, Dekker DM, *et al.* The emergence of reduced ciprofloxacin susceptibility in salmonella enterica causing bloodstream infections in rural Ghana. *Clin Infect Dis* 2016; **62**(suppl_1): S32–6.
64. Mercedes Berlanga M, Montero T, Hernández-Borrell J, *et al.* Influence of the cell wall on ciprofloxacin susceptibility in selected wild-type gram-negative and gram-positive bacteria. *Int J Antimicrob Agents* 2004; **23**(6): 627–30.
65. Barman Balfour JA, Wiseman LR. Moxifloxacin. *Drugs* 1999; **57**(3): 363–73.
66. Al Omari MMH, Jaafari DS, Al-Sou'od KA, *et al.* Moxifloxacin hydrochloride. *Profiles Drug Substances, Excipients Related Methodology* 2014; **39**:299–431.
67. Dalhoff A, Bowker K, MacGowan A. Comparative evaluation of eight in vitro pharmacodynamic models of infection: activity of moxifloxacin against escherichia coli and streptococcus pneumoniae as an exemplary example. *Int J Antimicrob Agents* 2020; **55**(1): 105809.
68. Ince D, Zhang X, Hooper DC. Activity of and resistance to moxifloxacin in staphylococcus aureus. *Antimicrob Agents Chemother* 2003; **47**(4): 1410–5.
69. Dubois J, Dubois M. Levonadifloxacin (wck 771) exerts potent intracellular activity against staphylococcus aureus in thp-1 monocytes at clinically relevant concentrations. *J Med Microbiol* 2019; **68**(12): 1716–22.