# HKFGCN: Predicting microbe-drug associations based on graph convolutional network with the multiple heterogeneous information kernel fusion

Ziyu Wu1, Shasha Li2, Lingyun Luo1,3, Pingjian Ding1

1. School of Computer Science, University of South China, Hengyang 421001, Hunan, China

2. Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong 999077, China

3. Hunan Medical Big Data International Sci.&Tech. Innovation Cooperation Base, Hengyang 421000, Hunan, China

# Abstract

Accumulating clinical studies have shown that the microbes in the human body closely interact with the human host and participate in regulating the energy efficiency of drugs. Identifying the link between microbes and drugs can facilitate the development of drug discovery and reuse, so microbes have become a new target for antimicrobial drug development. However, most of the links between microbes and drugs were discovered by biological experiments, which are time-consuming, expensive, and sometimes risky. Therefore, it is necessary to leverage computational ways to predict microbe-drug associations to aid biological experiments. In this study, we propose a new method, called HKFGCN (Heterogeneous information Kernel Fusion Graph Convolution Network), to predict the microbe-drug associations. Instead of mixing different topological information together, HKFGCN extracts different features of topological information separately, and further extracts Gaussian kernel features after extracting features. HKFGCN consists of three main steps. Firstly, we constructed two similarity networks and a microbe drug association network based on numerous biological data, including drug similarity and microbe similarity. Secondly, we used two kinds of encoders to extract features from the two networks. Based on the extracted features, we further obtained Gaussian kernel features from each layer of drug and microbe, and fused them. Finally, we reconstructed the bipartite microbe-drug graph based on the learned representations. Experiments show that the HKFGCN model performs well on the three datasets through three cross-validation methods. In addition, case studies were conducted on HIV and SARS-CoV-2, and the results were confirmed by existing literature.

# 1 Introduction

The human microbiome is a complex community of trillions of microbes, including microbe eukaryotes such as bacteria, archaea, viruses and fungi, protozoa and worms. These microbes exists in the metabolic process and the regulatory process of the immune system, ~~and they~~ affect the general health and homeostasis of the human body by actively participating in the activities of human body([1](#_ENREF_1)). People refer to microbes as "forgotten" organs because they improve metabolism, resist pathogens, synthesize essential vitamins and boost immunity while maintaining homeostasis in the human internal environment([2](#_ENREF_2)). However, an unbalanced microbiome may lead to diseases([3](#_ENREF_3)). For example, Type 1 diabetes (T1D), a debilitating autoimmune disease, is caused by T cell-mediated destruction of insulin-producing B cells. Interactions between gut microbes and the innate immune system are key epigenetic factors that alter susceptibility to T1D ([4](#_ENREF_4)). In addition, studies have shown that microbes can regulate drug efficacy and toxicity by participating in drug absorption and metabolism([5](#_ENREF_5)). For example, tenofovir can be efficiently degraded by the vaginal microbiome before it is converted into a drug-active form by host cells([6](#_ENREF_6)). On the other hand, the diversity and function of microbe communities living in the body are in turn altered by drugs. In the aspect of drug therapy, the human organism has been regarded as a new drug target([7](#_ENREF_7)). Therefore, predicting microbe-drug associations is beneficial and necessary for the treatment of microbe-based diseases as well as for the discovery of microbe-based drugs. However, traditional wet laboratory experiments to uncover microbe-drug associations are time consuming, expensive, and laborious. As an adjunct to biological experiments, computational methods can be used to predict potential microbe-drug associations and thus improve the efficiency of drug development.

In the past period,several databases have published scores of experimentally-proven associations between microbes and drugs, such as MDAD([8](#_ENREF_8)), aBiofilm([9](#_ENREF_9)) and DrugVirus([10](#_ENREF_10)). These free datasets have led to advances in deep learning-based discoveries of microbe-drug associations. For example, Deng et al. proposed a framework Graph2MDA([11](#_ENREF_11)) based on variational graph autoencoder (VGAE) to predict associations between microbes and drugs. Multiple Kernel Learning (MKL) is often used to improve the prediction performance on bipartite networks. Therefore, it is not suitable to use multi-kernel learning on samples with few feature types. In order to make up for the lack of preconditions in the use of multiple kernel learning, Yang et al. proposed a framework MKGCN([12](#_ENREF_12)) ,  which regarded the features obtained from each GCN layer as different feature types, because different GCN layer represented different structural information of nodes. In addition, the link between microbes and drugs can also be examined using existing models of bipartite biological networks. For example, LAGCN([13](#_ENREF_13)), is a bipartite drug-disease prediction model based on convolutional neural networks, which was originally used to predict drug-disease associations. However, this kind of models often build an integrated network by mixing different topological information together, without distinguishing their domains (i.e. microbe and drug domains), which may lead to the loss of a large amount of network-specific information.

In this study, a new fusion model, HKFGCN, is proposed to predict the associations between microbes and drugs. It is based on Gaussian kernel fusion graph convolution network with heterogeneous information. We constructed three networks, namely, drug-drug similarity network, microbe-microbe similarity network, and observed microbe-drug association network. Also, we designed two different feature extraction modules for different network topologies. The first feature extraction module is the GCN([14](#_ENREF_14)) module composed of a multi-layer connected neural network architecture, which is used to learn the low-dimensional representation of nodes from the graph structures. In this study, we use it to extract the features of microbe and drug similarity networks respectively. The second feature extraction module is the BGCN aggregator([15](#_ENREF_15)), which consists of a bilinear aggregator and a traditional GCN aggregator. We use it to extract the interactive features between microbes and drugs. The features extracted from the two modules were fused to obtain the final embeddings of drugs and microbes. Then, we treat the embeddings in each layer as different features to obtain the multi-kernel matrix, and fuse the matrix using the average weighting method. Finally, new microbe-drug associations are predicted by joint kernels in microbe-drug space using dual graph-regularized least squares (DLapRLS) ([16](#_ENREF_16)).

# 2 Materials and Methods

This section describes the model HKFGCN details. The workflow of HKFGCN for predicting potential microbe and drug associations is shown in Figure 1. HKFGCN consists of three main steps. Firstly, two similarity networks containing rich biological and a microbe-drug association network were constructed, including drug similarity and microbe similarity. Secondly, we used two kinds of encoders to extract features from different networks, based on the extracted features, we further obtained the Gaussian kernel features of each layer of drug and microbe, and fused them. Finally, we reconstructed the bipartite microbe-drug associations based on the learned representation.

Graphical user interface

Description automatically generated

Figure The workflow of HKFGCN. a) Two similarity networks and a microbe-drug association network based on rich biological were constructed, including drug similarity and microbe similarity. b) Two kinds of encoders were used to extract features from different networks. Based on the extracted features, Gaussian kernel features of each layer of drug and microbe were further obtained and fused. c) The bipartite microbe-drug association was reconstructed based on the learned representations.

## 2.1 Datasets

We evaluated the performance of HKFGCN on three different datasets: MDAD([8](#_ENREF_8)), aBiofilm([9](#_ENREF_9)) and DrugVirus([10](#_ENREF_10)). The MDAD dataset (<http://chengroup.cu-mt.edu.cn/MDAD/>) includes 5055 clinical or experimental data for 80 microbes and 1388 drugs. After removing redundant, we had a total of 2,470 associations involving 1373 drugs and 173 microbes. Unique anti-biofilm drugs for organisms including bacteria and fungus were recorded in the aBiofilm dataset ([http://bioinfo.imtech.res.in/ manojk/abiofilm/](http://bioinfo.imtech.res.in/%20manojk/abiofilm/)). We ended up selecting 2884 microbe-drug associations involving 1720 drugs and 140 microbes. The last is the DrugVirus dataset (<https://drugvirus.info/tech_doc/>), which contains the development of various human viral activities and related compounds including the novel coronavirus SARS-COV-2. In the end, we screened 933 associations from 95 viruses and 175 drugs. In summary, Table 1 shows the detailed and their sizes for the three datasets described above.

Table : Statistics for three data sets.

|  |  |  |  |
| --- | --- | --- | --- |
| Datasets | Microbes | Drugs | Associations |
| MDAD | 173 | 1373 | 2470 |
| aBiofilm | 140 | 1720 | 2884 |
| DrugVirus | 95 | 175 | 933 |

## 2.2 Network Construction

### 2.2.1 Microbe-drug association network

We represent the known microbe-drug association network as graph G, which contains N drugs and M microbes, and its adjacent matrix was . if a drug is associated with a microbe . On the contrary, indicates that the association is not known.

### 2.2.2 Drug–drug similarity network

The drug-drug similarity network containing N drugs is represented by graph , and the drug similarity matrix constitutes its adjacency matrix . Specifically, if drug is the first k-nearest (top k) neighbor of drug evaluated by the drug similarity matrix , then the entry of is ; otherwise . Drug similarity consists of two parts, which are the combination of drug structure similarity and drug Gaussian kernel similarity.

We used SIMCOMP2([17](#_ENREF_17)) to calculate the structural similarity of drugs and was used to represent the structural similarity matrix of drugs. Another similarity measure was calculated using the Gaussian kernel similarity of the drug, which was obtained on the basis of the drug structural similarity matrix to complement the drug similarity([18](#_ENREF_18)). Gaussian kernel similarity of drugs implies that drugs with similar efficacy interact with similar microbes. indicates that it is in the *i-*th row of the adjacency matrix of drug structural similarity as interaction profiles for drug , and the Gaussian nuclear similarity between drug and drug was expressed as Equation 1:

where is defined as Equation 2 and represents the normalized kernel bandwidth([19](#_ENREF_19)):

where is the original bandwidth.

We combined drug structure similarity DS and drug Gaussian kernel similarity GD to obtain comprehensive drug similarity . The new drug similarity degree was calculated as Equation 3:

### 2.2.3 Microbe–microbe similarity network

Similarly, for microbe-microbe similarity network, we used graph to represent microbes in the microbe-microbe similarity network, and the microbe similarity matrix is composed of its adjacency matrix .  Specifically, if microbe is the first k-nearest (top k) neighbor of microbe evaluated by the drug similarity matrix , then the entry of is ; otherwise . The construction of microbe similarity is similar to the construction of drug similarity, which combines the functional similarity of microbes with the Gaussian kernel similarity.

We used the Kamneva tool([20](#_ENREF_20)) to calculate the functional similarity of microbes and was used to represent the structural similarity matrix of microbes. Similarly, We use Gaussian kernel similarity of microbes to calculate another similarity measure to supplement microbe similarity([18](#_ENREF_18)). indicates that it is in the *i-*th column of the adjacency matrix of microbe functional similarity as interaction profiles for microbe , and the Gaussian nuclear similarity between microbe and microbe was expressed as Equation 4:

is defined as Equation 5, which represents the normalized kernel bandwidth ([19](#_ENREF_19)):

Where is the original bandwidth.

Similarly, drug similarity is composed of two parts: microbe functional similarity MF and microbe Gaussian kernel similarity GM, which are fused to obtain drug similarity . The new microbe similarity degree was calculated as Equation 6:

### 2.3 Encoder

Two encoders are used here, one is to extract features from two similarity networks (microbe similarity network and drug similarity network), and another is used to extract features between microbe and drug association network. The first encoder uses GCN to extract features of microbe networks and drug networks respectively. The second encoder is used to extract features for the microbe-drug association network, which consists of a bilinear aggregator (BA) and a traditional GCN aggregator (AGG).

**2.3.1 Encoder for** **similarity networks based on GCN.**

GCN([14](#_ENREF_14)) is a multi-layer connected neural network architecture, which is widely used in the aggregation of node features in networks. The essence of graph convolution on network is to achieve feature aggregation among related nodes. Here, the GCN encoder mainly extracts the embeddings of drugs and microbes respectively based on the drug-drug and microbe- microbe similarity networks.

First, we defined the initial drug and microbe embeddings as follows:

Therefore, the embeddings obtained from two similar networks via the GCN aggregator are defined as follows:

where is the drug output embeddings at the *l-*th -layer, is the microbe output embeddings at the *l-*th-layer, is the representation of the drug input embedded in layer L, is the representation of the microbe input embedded in layer L. and are the *l-*th layer feature extraction modules, and they are trainable matrices. The convolution operation of the graph is expressed as , and its calculation is as Equation 9:

Whereis a ReLU([21](#_ENREF_21)) activation function and .

**2.3.2 Encoder for association networks based on BGCN.**

BGCN is a new graph convolution operator([15](#_ENREF_15)) with the properties of permutation invariance and linear computational complexity. It enhances weighted sums by pairwise interactions represented by neighborhood nodes. Specifically, for a drug , it’s drug feature extraction module for microbe-drug association network is defined as:

where is an element-wise product, is the *l-*th drug feature of drug in microbe-drug association network, is a trainable matrix and is the weighting factor that balances the effects of the first and second terms. is adjacent matrix of microbe-drug association network. Similarly, for a microbe , it’s microbe feature extraction module for microbe-drug association network is defined as:

is the *l-*th microbe feature of microbe in microbe-drug association network.

The feature of the associated network is as follows:

we merged similarity networks features and association networks features as follows:

### 2.4 Combined kernel on embedding

The embedding of each layer of the encoder aggregates the information of different neighbors of the node. The original feature of a node in a heterogeneous graph is represented by ,  and H1 is the feature obtained by aggregating the first-order neighbor information of nodes on the basis of H0.  Similar to MKGCN([12](#_ENREF_12)), we treat the embeddings of each layer as different eigenvectors and calculate them to obtain multiple kernel matrices.

For each layer of embedding, we can divide it into drug embedding and microbe embedding, as shown in Equation 13. For the nuclear matrix of drug and microbe embedded in each layer, we used Gaussian interaction profile (GIP) to calculate as follows:

where and are drug and microbe embedding profiles at the *i-*th row of the layer *l*; denotes the corresponding bandwidth.

The kernel set of drug space and microbe space are obtained by combining the existing similarity matrices, where is drug similarity matrix and is microbe similarity matrix.

To fuse the above kernel (in two Spaces) and improve the reliability of predicting microbe-drug associations, we used a weighted sum approach. A composite kernel can be defined as:

where and are the *i-*th kernel in the drug and the *i-*th kernel in the microbe, respectively. The weight of each kernel is represented by  and . Here, we set .

### 2.5 Decoder and Optimization

In order to better reconstruct the connection between drugs and microbes, our decoder is formulated as follows:

where is the predicted probability score matrix. The prediction score is given by , where the corresponding term is the association between drug d*i* and microbe . and are trainable matrices.

Dual Laplacian Regularized Least Squares (DLapRLS)([16](#_ENREF_16))  is a kernel matrix model based on two feature Spaces, which increases the regularization function of graphs and improves the prediction ability. Therefore, we apply the DLapRLS framework to predict microbiological drug associations. We define the loss function as follows:

Which and  are the fusion kernel of drug and microbe respectively, and is the adjacency matrix for microbe-drug associations in the training set,is the Frobenius norm. and A and B are defined as follows, which is the normalized Laplacian matrices:

where and are diagonal matrices as follow:

There are two parameters that need to be optimized in our model. The first is about the GCN parameter, based on previous studies([22](#_ENREF_22)), In order to minimize the loss function, we use Adam([23](#_ENREF_23)) optimizer to optimize the model as described in ([24](#_ENREF_24)) . The second parameter is about DLapRLS. By calculating the partial derivative, we can directly obtain the iterative function. We update DLapRLS parameters with Equation 24 and Equation 25, respectively.

# 3 Result

In this section, we first briefly describe the setting of our experimental parameters, and then demonstrate the performance of our HKFGCN model by comparing with three existing methods and verify the effectiveness of the model by ablation experiments. Finally, we selected some microbes and drugs to make predictions for the case study.

A common method for evaluating the effectiveness of a prediction is k-fold cross validation (K-CV). To compare with other existing tools on three datasets (MDAD, aBiofilm, and drug viruses), we used three cross-validation methods (2-CV, 5-CV, and 10-CV) on HKFGCN. Taking 5-CV as an example, 20% of known microbe-drug association pairs and 20% of unknown microbe-drug association pairs were randomly selected from the dataset as the test set. The remaining 80% of clinically reported microbe-drug association pairs and unknown microbe-drug association pairs were used for training models. We used two indicators for evaluation: the area under the receiver operating characteristic curve (AUC)([25](#_ENREF_25)) and the area under the precision recall curve (AUPR)([26](#_ENREF_26)).

Our proposed HKFGCN framework adopts a three-layer architecture with 64 hidden units per layer and sets the learning rate in the optimization algorithm to 0.01 and epoch=20.

## 3.1. Parameter selection

There are several hyperparameters in the model, such as , and the number of neighbors k of microbes and drugs. During parameter selection, under 5-CV, we performed all evaluation experiments on the MDAD dataset. In hyperparameter selection, we take into account all possible combinations of parameter values, as follows: for k and for , .

First, we studied the effect of the parameter k, which represents the number of k nearest neighbors for each drug or each disease. Therefore, we need to choose the value of k that makes the best performance of the model. We choose the final value of k by changing the value of k to observe the AUPR of the result. The step of k (from 1 to 20) is 1. Fig2 shows the AUPR of different k, showing the effect of k on the predictive performance. The results show that the best result is obtained when the value of k is 4, but we chose the value of k as 10 to make the model more stable.

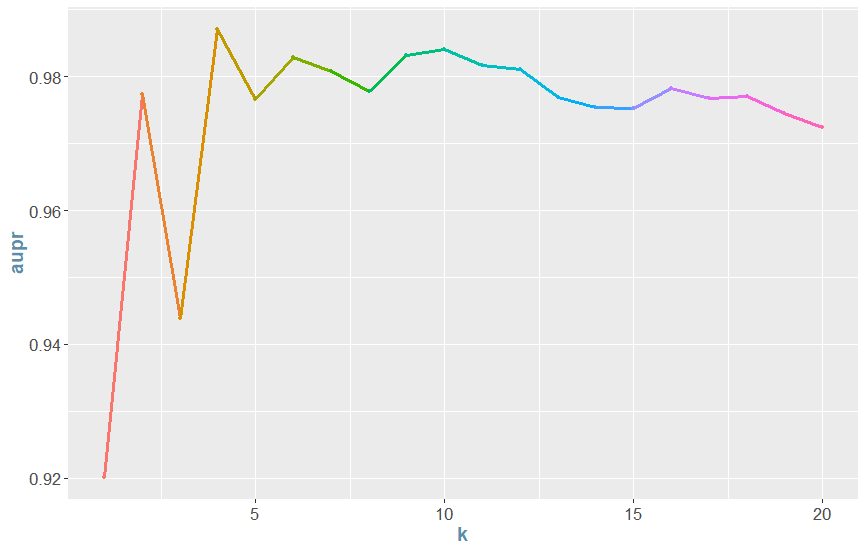


Figure 2 AUPR for different values of K on dataset MDAD.

In the objective function Equation (19), and ,  which represent the weight of the graph regularization term in DLapRLS, and we need to study these two parameters because they are important parameters of the model. Figure 3 shows that the value of AUPR varies with different and . It can be seen that the model is relatively effective when and . It can be concluded that the model works well when the difference between and is small.

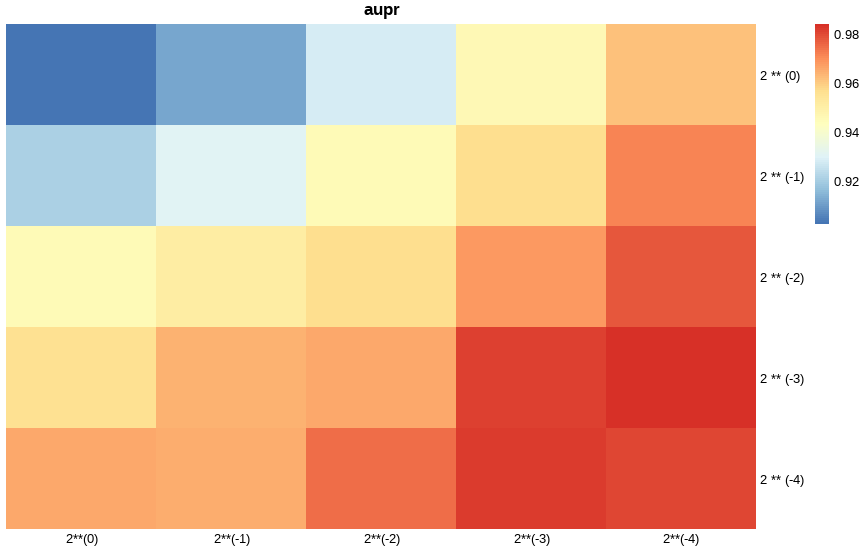


Figure shows that the value of AUPR varies with different λd and λm.

## 3.2 Comparison with existing methods

To evaluate the performance of our proposed model, we compare HKFGCN with three state-of-the-art approaches listed below. As mentioned earlier, few methods have been proposed for microbe-drug association prediction. Therefore, some of the following approaches are for other link prediction tasks in computational biology, such as LAGCN.

**Graph2MDA**: Models for predicting microbe-drug associations are based on variational graph autoencoders (VGAE).

**MKGCN**: This paper proposes the application of multiple kernel learning (MKL) to the prediction of microbe-drug association.

**LAGCN**: is a convolutional neural network-based model for predicting drug-disease associations.

To be fair, we used the default parameter values of the original implementation for each of the three existing methods, and we also compared them on the same benchmark MDAD([8](#_ENREF_8)), aBiofilm([9](#_ENREF_9)) and DrugVirus([10](#_ENREF_10))datasets.

As shown in Figure 4, the AUC value of our HKFGCN model is higher than that of the other models on the three datasets under the condition of 5-CV. The AUC value of our model on dataset MDAD is 0.9995, which is 1.14% higher than that of the suboptimal method MKGCN. On the dataset aBiofilm, the AUC value of the second-best method MKGCN is 0.9939, and the AUC value of our model is 0.9996, which is 0.5% higher than MKGCN. On dataset DrugVirus, the AUC value of our model was 0.9861, and the second-best method MKGCN was 0.9814. In another indicator AUPR, for the datasets MDAD and aBiofilm, the AUPR of our model is slightly higher than that of other models, which are 0.9752 and 0.9849 respectively, and the second-best method is 0.9673 and 0.9828. On the dataset DrugVirus, the AUPR of our model is slightly lower than that of the other models. In general, our model achieves better results than other models on the three datasets.

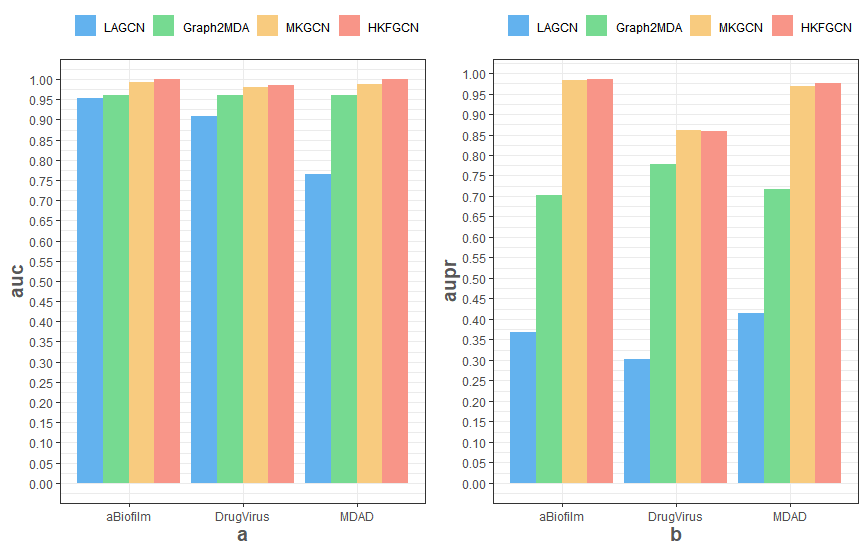


Figure 4 Performance comparisons of HKFGCN with three competitive methods on three datasets. (a) Results of AUC for different models on three datasets. (b) Results of AUPR for different models on three datasets.

## 3.3 ablation study

Our HKFGCN model consists of three main components, including 1) we use different encoders on the association network and the similarity network. 2) On the basis of the features obtained by the encoder, we add Gaussian kernel. 3) Attention mechanism is added in the process of Gaussian kernel fusion, as shown in Figure 1. Here, we performed ablation studies in the 5-fold CV case using the dataset MDAD to assess the impact of each component on the model. Our variable model for ablation experiments was as follows.

**HKFGCN w/o Bgcn:** it has no different encoder from similarity network in feature extraction of association network.

**HKFGCN w/o Kernel:** it has no Gaussian kernel after feature extraction by encoder.;

**HKFGCN w/o Attention:** equal weight is used in Gaussian kernel instead of bias weight;

Table Comparison analysis between HKFGCN and its Variants on MDAD dataset.

|  |  |  |
| --- | --- | --- |
|  | AUC | AUPR |
| HKFGCN w/o Kernel | 0.9722 | 0.6195 |
| HKFGCN w/o Bgcn | 0.9990 | 0.9669 |
| HKFGCN w/o attention | 0.9994 | 0.9719 |
| HKFGCN | 0.9995 | 0.9752 |

Table 2 shows the comparison of AUC and AUPR on HKFGCN and its three variants. We observe that both Gaussian Kernel and different encoders play an important role in HKFGCN, because both HKFGCN w/o Kernel and HKFGCN w/o Bgcn have a great influence on the results, among which HKFGCN w/o Kernel has the greatest influence. Because HKFGCN w/o Kernel has the lowest performance, its AUC is 0.9722, AUPR is 0.6195, followed by HKFGCN w/o Bgcn. Secondly, BGCN has the greatest impact on HKFGCN, with AUC of 0.9990 and AUPR of 0.9669. This indicates that adding different encoders for different networks is helpful for feature extraction. Note that the effect of the mechanism on the HKFGCN model is relatively small, but it is also helpful to improve the performance of the model. In general, these three parts are important components of HKFGCN.

## 3.4 Case study

To further validate the prediction performance of HKFGCN, we selected two prediction targets from the MDAD dataset and DrugVirus dataset for case study. The MDAD dataset targets human immunodeficiency virus (HIV). The DrugVirus dataset targets SARS-CoV-2, a virus that can cause respiratory, intestinal, liver, and nervous system diseases in its host.

For each selected prediction target, we treat its corresponding known items as unknown, and sort their prediction results in descending order according to the scores. We measured the performance of our model by examining and collating previous reports to verify that the most likely drug candidates appeared in the top 20 and 50 lists.

AIDS is an infectious disease caused by HIV, which is devastating to the human body. It takes CD4T lymphocytes as the main target and is the most important cell in the human immune system, making the human body prone to various diseases, malignant tumors and high mortality. Of the top 20 drugs predicted by case studies for Human Immunodeficiency Virus (HIV), 95% were supported by the literature. Table 3 shows the published drug names and PMIDs, such as Lopinavir (LPV), a new protease inhibitor against HIV, which significantly improved the pharmacokinetic properties of Lopinavir when administered in combination with low-dose ritonavir. Thus, the anti-HIV-1 protease activity was increased.

Table The top 20 predicted human immunodeficiency virus (HIV) drugs. Columns one to three record the top 10 drugs, and columns four to six record the top 11 to 20 drugs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | Drug | Evidence | Rank | Drug | Evidence |
| 1 | Lopinavir | PMC2621403 | **11** | Didanosine | PMID: 10651392 |
| 2 | Saquinavir | PMID: 7715294 | **12** | Delavirdine | PMID: 11152019 |
| 3 | Atazanavir | PMID: 19053892 | **13** | Ceftazidime | PMID: 11527042 |
| 4 | Indinavir | PMID: 10651394 | **14** | Stavudine | PMID: 10595868 |
| 5 | P60.4Ac | Unconfirmed | **15** | Bevirimat | PMID: 19024627 |
| 6 | Tipranavir | PMID: 18578560 | **16** | Rilpivirine | PMID: 26049950 |
| 7 | Darunavir | PMID: 19323590 | **17** | Amprenavir | PMID: 11152018 |
| 8 | Nelfinavir | PMID: 10776836 | **18** | Brecanavir | PMID: 17491001 |
| 9 | Lamivudine | PMID: 20001611 | **19** | Dolutegravir | PMID: 24982751 |
| 10 | Epigallocatechin Gallate | PMID: 21730371 | **20** | Enfuvirtide | PMID: 15110129 |

The novel coronavirus infection (SARS-CoV-2) is very serious and has affected global health. There is an urgent need to find effective preventive and therapeutic drugs. Therefore, we selected SARS-CoV-2 as one of the case studies to test the predictive power of our model by predicting potentially effective therapeutic agents for SARS-CoV-2. Table 4 lists the top 50 predicted drugs associated with SARS-CoV-2. Of the top 50 drugs predicted to target human immunodeficiency virus (HIV), 90% are supported by the literature. For example, remdesivir (RDV) effectively inhibited SARS-CoV-2 replication in human lung cells and primary human airway epithelial cell cultures (EC50 = 0.01 μm).

These predictions indicate that the HKFGCN model can predict potential associations in microbe-drug networks.

Table Top 50 prediction results obtained with SARS-COV-2 as the prediction target.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | Drug | Evidence | Rank | Drug | Evidence |
| 1 | Remdesivir | PMID: 32668216 | **26** | Azithromycin | PMID: 32853038 |
| 2 | Chloroquine | PMID: 32306288 | **27** | Chlorpromazine | PMID: 32387014 |
| 3 | Ribavirin | PMID: 34302258 | **28** | Amiodarone | PMID: 32737841 |
| 4 | Favipiravir | PMID: 35430987 | **29** | Suramin | PMID: 32513797 |
| 5 | Amodiaquine | PMID: 32486229 | **30** | ABT-263 | Unconfirmed |
| 6 | BCX4430 (Galidesivir) | PMID: 34551346 | **31** | Raloxifene | PMID: 35614039 |
| 7 | Nitazoxanide | PMID: 33031085 | **32** | Cyclosporine | PMID: 33606519 |
| 8 | Mycophenolic acid | PMID: 32579258 | **33** | Nelfinavir | PMID: 33910617 |
| 9 | Gemcitabine | PMID: 32432977 | **34** | Anisomycin | PMID: 35818229 |
| 10 | EIPA (amiloride) | PMID: 33299997 | **35** | Ritonavir | PMID: 32344308 |
| 11 | Luteolin | PMID: 34236507 | **36** | Mefloquine | PMID: 35215969 |
| 12 | Niclosamide | PMID: 34482191 | **37** | Alisporivir | PMID: 32376613 |
| 13 | Emetine | PMID: 34254564 | **38** | Homoharringtonine | PMID: 32251767 |
| 14 | Brequinar | PMID: 33584854 | **39** | Glycyrrhizin | PMID: 33918301 |
| 15 | Obatoclax | PMID: 34989664 | **40** | Tilorone (Amixin) | PMID: 32215760 |
| 16 | Hydroxychloroquine | PMID: 33465165 | **41** | Dalbavancin | PMID: 33262453 |
| 17 | Lopinavir | PMID: 32589165 | **42** | Rapamycin (Sirolimus) | PMID: 33807743 |
| 18 | Silvestrol | PMID: 32374474 | **43** | Benztropine | Unconfirmed |
| 19 | Ivermectin | PMID: 33592050 | **44** | Oritavancin | Unconfirmed |
| 20 | Eflornithine | Unconfirmed | **45** | Teicoplanin | PMID: 33675235 |
| 21 | Cepharanthine | PMID: 33423067 | **46** | Tamoxifen | PMID: 34934049 |
| 22 | Arbidol (Umifenovir) | PMID: 32360231 | **47** | 4-HPR (Fenretinide) | PMID: 32471278 |
| 23 | Dasatinib | PMID: 33008453 | **48** | Itraconazole | PMID: 33666253 |
| 24 | Berberine | PMID: 33670363 | **49** | Emodin | PMID: 21356245 |
| 25 | Sunitinib | PMID: 33083006 | **50** | Sorafenib | Unconfirmed |

# 4 Discussion and conclusion

In this paper, we proposed a new Multiple Kernel Learning (MKL) based model to predict the correlation between microbes and drugs. This model uses different encoders for different networks and performs multiple kernel learning on this basis. Experimental results show that our proposed HKFGCN model outperforms the state-of-the-art methods. We believe that the main contribution of our work is in at least three aspects :1) In order to integrate the rich biological information, including microbe genetic, pharmacochemical information, and microbe-drug associations, we first constructed three networks. 2) We used different encoders to extract features of different networks, effectively integrating the rich similarity and ontology information of microbes and drugs. 3) We further extracted the features obtained by the encoder and calculate the multi-layer embedding algorithm to obtain multiple kernel matrices. HKFGCN experimentally indicates that our model has A good prediction effect due to its excellent performance on three existing microbe-drug associations datasets: MDAD, aBiofilm, and DrugVirus. In addition, case studies of Human Immunodeficiency virus (HIV) and SARS-COV-2 further confirmed the effectiveness of our model in identifying potential drugs in microbes.

Although we have achieved good results with different encoders for different networks, there is still room to improve our prediction model in terms of encoder selection, because our encoder for feature extraction of similar networks is only the traditional GCN encoder. In the future, we can try more feature extraction modules and select the ones that are more suitable for different networks to improve the prediction performance of our model.

# Reference

1. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front Microbiol. 2021;12:634511.

2. Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, et al. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat Rev Microbiol. 2009;7(1):61-71.

3. Whiteside SA, Razvi H, Dave S, Reid G, Burton JP. The microbiome of the urinary tract--a role beyond infection. Nat Rev Urol. 2015;12(2):81-90.

4. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, et al. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature. 2008;455(7216):1109-13.

5. Zimmermann M, Patil KR, Typas A, Maier L. Towards a mechanistic understanding of reciprocal drug-microbiome interactions. Mol Syst Biol. 2021;17(3):e10116.

6. Klatt NR, Cheu R, Birse K, Zevin AS, Perner M, Noël-Romas L, et al. Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. 2017;356(6341):938-45.

7. Kashyap PC, Chia N, Nelson H, Segal E, Elinav E. Microbiome at the Frontier of Personalized Medicine. Mayo Clin Proc. 2017;92(12):1855-64.

8. Sun YZ, Zhang DH, Cai SB, Ming Z, Li JQ, Chen X. MDAD: A Special Resource for Microbe-Drug Associations. Front Cell Infect Microbiol. 2018;8:424.

9. Rajput A, Thakur A, Sharma S, Kumar M. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. Nucleic Acids Res. 2018;46(D1):D894-D900.

10. Andersen PI, Ianevski A, Lysvand H, Vitkauskiene A, Oksenych V, Bjoras M, et al. Discovery and development of safe-in-man broad-spectrum antiviral agents. Int J Infect Dis. 2020;93:268-76.

11. Deng L, Huang Y, Liu X, Liu HJB. Graph2MDA: a multi-modal variational graph embedding model for predicting microbe–drug associations. 2022;38(4):1118-25.

12. Yang H, Ding Y, Tang J, Guo F. Inferring human microbe–drug associations via multiple kernel fusion on graph neural network. Knowledge-Based Systems. 2022;238.

13. Han H, Zhu R, Liu JX, Dai LY. Predicting miRNA-disease associations via layer attention graph convolutional network model. BMC Med Inform Decis Mak. 2022;22(1):69.

14. Kipf TN, Welling MJapa. Semi-supervised classification with graph convolutional networks. 2016.

15. Zhu H, Feng F, He X, Wang X, Li Y, Zheng K, et al. Bilinear graph neural network with neighbor interactions. 2020.

16. Ding Y, Tang J, Guo F. Identification of Drug–Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. Knowledge-Based Systems. 2020;204.

17. Hattori M, Tanaka N, Kanehisa M, Goto S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res. 2010;38(Web Server issue):W652-6.

18. Long Y, Wu M, Kwoh CK, Luo J, Li X. Predicting human microbe-drug associations via graph convolutional network with conditional random field. Bioinformatics. 2020;36(19):4918-27.

19. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011;27(21):3036-43.

20. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. PLoS Comput Biol. 2017;13(2):e1005366.

21. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. Icml; 2010.

22. Yu Z, Huang F, Zhao X, Xiao W, Zhang W. Predicting drug-disease associations through layer attention graph convolutional network. Brief Bioinform. 2021;22(4).

23. Kingma DP, Ba JJapa. Adam: A method for stochastic optimization. 2014.

24. Glorot X, Bengio Y, editors. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics; 2010: JMLR Workshop and Conference Proceedings.

25. Hand DJ, Anagnostopoulos C. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters. 2013;34(5):492-5.

26. Boyd K, Eng KH, Page CD, editors. Area under the precision-recall curve: point estimates and confidence intervals. Joint European conference on machine learning and knowledge discovery in databases; 2013: Springer.