Drug Repositioning based on Graph Convolutional Network

Xinyi Zhang[1], **Yanni Xu** [1], **Yuling Lin** [1], **Menglong Zhang** [1]

[1] 23020221154148, [1] 23020221154133, [1] 23020221154100, [1] 23020221154145

Abstract

Drug discovery is characterized by time-consuming, high cost and high risk, while drug repositioning provides a relatively low-cost and efficient method to rapidly discover effective treatments. Drug-disease association prediction is a key foundation for drug repositioning, and it is an important task to develop a more effective prediction method in the face of limitations such as the misfit of existing model architectures and data features. In this paper, a Multiple Kernel fusion model based on Graph Convolutional Network (called MKGCN) is developed for predicting drug-disease associations. The model first integrates known drug-disease associations and drug similarity and disease similarity information to construct a heterogeneous network, then extracts multilayer features by Graph convolutional network (GCN). Then, we uses the embedding features of each layer to calculate the kernel matrix, and uses the average weighting method to fuse the kernel matrix. Finally, Dual Laplacian Regularized Least Squares (DLapRLS) is applied to predict new drug-disease associations. A 5-fold cross-validation is performed on the drug-disease association dataset from CTD (Comparative Toxicogenomics Database), and the experimental results show that MKGCN is a more effective prediction method with significantly higher AUPR values than existing state-of-the-art prediction methods, while it is also adaptable on other datasets. In addition, the case study also confirmed the reliability of the predicted novel drug-disease associations in terms of biological explanatory aspects.

Introduction

Currently, the research and development (R&D) of new drugs has the characteristics of long cycle, high cost and high risk (Scannell et al. 2012). At the same time, it also has a high attrition rate, with most drug developments ending in failure due to high toxicity or ineffectiveness (Plenge, Scolnick, and Altshuler 2013). Despite rapid technological advances and significant increases in R&D costs, the number of new drugs brought to market each year is decreasing (Booth and Zemmel 2004). In this context, Ashburn and Thor first proposed the "drug repositioning" approach in 2004 to explore new indications and targets for drugs already on the market (Plenge, Scolnick, and Altshuler 2013).Due to the significant saving of a lot of development time and cost, drug repositioning has attracted considerable attention as it has addressed a development pain point in the pharmaceutical industry. The main challenge in drug repositioning is to detect the molecular target of a drug among hundreds to thousands of gene products that have indirect responses to changes in target activity. However, traditional statistical analysis methods are unable to discover the molecular targets of drugs from the large number of genes.Graph Convolutional Network (GCN) (Kipf and Welling 2016), as a generalization of convolutional networks on graph structure, has performed well in the field of biomedical network analysis, such as miRNA-drug resistance association prediction (Huang et al. 2020), multiple drug side effect prediction (Zitnik, Agrawal, and Leskovec 2018), miRNA-disease association prediction (Li et al. 2020), etc.With this background, we will propose a drug-disease association prediction model based on graph convolutional networks.

Related Work

Research status of drug repositioning.

In recent years, advances in bioassay technology have yielded a large amount of data, such as information on drug chemical structures, drug-targeting proteins, disease drug associations and so on. The large publicly available databases provide a reliable source of data for drug repositioning studies and drive the development of relevant predictive models. The first is drug repositioning approaches based on text mining and semantic reasoning. For example, Chen et al. (Chen, Ding, and Wild 2012) proposed a method based on semantic connection networks to predict drug-target associations. The next approach is the method of using network models, such as Wu et al. (Wu et al. 2013), who used the drug-disease heterogeneous network model to identify the close connection module between drugs and diseases, thereby extracting potential drug-disease association information for drug repositioning.Finally, machine learning technology has been widely used to develop more accurate drug disease association prediction models. Kai Yang et al. proposed a method called HED to predict potential associations between drugs and diseases based on the drug-disease heterogeneous network (Yang et al. 2019).

Research status of deep learning.

Deep learning is an important branch of machine learning, which has been widely used in natural language processing, computer vision, reinforcement learning and other aspects. Its advantage is that it can mine complex structural relationship information between input features and output decisions from large-scale data. The application of deep learning in drug discovery and molecular informatics is still in its infancy, but it has shown great potential. Aliper and Plis use deep neural networks (DNN) to analyze gene expression profile data to predict treatment categories of drugs (Aliper et al. 2016). The experimental results show that the classification accuracy of DNN exceeds that of SVM, which indicates that deep learning is a useful tool in the field of drug development.Hu et al. (Hu et al. 2019) introduced a convolutional neural net-work model to unveil drug-target interactions.Zeng et al. (Zeng et al. 2019) used multi-modal deep autoencoder and variational autoencoder models to discover drug-disease associations.

Research status of GCN.

As a special deep neural network structure, Convolutional Neural Network (CNN) can effectively reduce the complexity of traditional neural networks and is mainly used to process structured data such as images. Due to unsatisfied dimensional consistency and sequence order, the traditional convolution on grid structures cannot be applied directly to graphs. By studying the generalization of convolution operators on non-Euclidean structured data, a graph convolution network (GCN)(Kipf and Welling 2016) adapted to graph structures was finally generated. The graph convolutional network captures structural information of the graph through messages passing between the nodes of the graph and maintains high interpretability. It has shown convincing performance in biomedical network analysis, such as microRNA (miRNA)-disease association prediction (Huang et al. 2020), multidrug side effect prediction (Zitnik, Agrawal, and Leskovec 2018) and miRNA-drug resistance association prediction (Li et al. 2020).

Method

Problem Description

In this paper, diseases and drugs are represented as two kinds of nodes in the network, the node set of N_r drugs is denoted by $R = \{r_1, r_2, ..., r_{N_r}\}$, and the node set of N_d disease is denoted by $D = \{d_1, d_2, ..., d_{N_d}\}$. The edges in the network represent the relationship between the disease and the drug, which is represented by the adjacency matrix $Y \in \mathbf{R}^{N_r \times N_d}$, where $Y_{ij} = 1$ means that the drug is related to the disease, and $Y_{ij} = 0$ means that the relationship is unknown. The model prediction task is to obtain a prediction matrix $\mathbf{F}^*$ of the same size as Y for discovering new disease-drug associations.

Model Architecture

We build a graph convolutional network-based multi-kernel fusion model (MKGCN) for predicting drug-disease associations. The model is based on the heterogeneous network of diseases and drugs, extracts multi-layer embedded features through GCN, calculates the kernel matrix through the embedded features of each layer and performs weighted fusion, and finally applies the dual Laplacian regularized least square method (DLapRLS) to predict potential the drug-disease association, the algorithm flow chart is shown in Figure 1.

Drug	Disease	Association	Sparsity
269	598	18416	0.1145

Table 1: Statistics for the dataset

Construction of the heterogeneous network

The data set used in this article comes from the research of Wen (Zhang et al. 2018a). The basic information of the data set is shown in Table 1, which contains 18,416 disease-drug associations between 269 drugs and 598 diseases. The association information comes from the CTD database (Davis et al. 2017). In this paper, the drug-disease association matrix is defined as $Y \in \mathbf{R}^{N_r \times N_d}$.

This paper mainly calculates the similarity based on the target protein characteristics of the drug, and combines it with the drug Gaussian kernel similarity to construct the drug similarity kernel.

Firstly, the target protein similarity of the drug is calculated, and the drug is encoded into a binary feature vector, where each element indicates whether the corresponding feature exists. In this paper, the Jaccard index (Zeng et al. 2019) is used to calculate the target protein similarity of the drug, and the calculation formula is as follows:

$$S_{ij}^r = \frac{|x_i \cap x_j|}{|x_i \cup x_j|}, \tag{1}$$

where $|x_i \cap x_j|$ represents the number of times that the elements in x_i and the corresponding elements in x_j are both equal to 1, and $|x_i \cup x_j|$ represents the number of times that the elements in x_i or the corresponding elements in x_j are equal to 1.

The Gaussian interaction profile kernel similarity between drugs is then calculated, which is used to supplement the missing term of target protein similarity.

$$GR\left(r_i, r_j\right) = \exp\left(-\eta_r \left\|Y_{r_i} - Y_{r_j}\right\|^2\right), \tag{2}$$

where y_{d_i} represents the i-th row in the adjacency matrix, that is, the interaction spectrum of the drug r_i, and η_r represents the normalized nuclear bandwidth (Van Laarhoven, Nabuurs, and Marchiori 2011), defined as follows:

$$\eta_r = \frac{\eta_r'}{\left(\frac{1}{N_r}\sum_{i=1}^{N_r}\left\|Y_{r_i}\right\|^2\right)}, \tag{3}$$

where η_r is the raw bandwidth.

The comprehensive drug similarity $\mathbf{K}_s^r$ of two drugs r_i, r_j is defined as follows:

$$\mathbf{K}_s^r\left(r_i, r_j\right) = \begin{cases} \frac{S_{ij}^r + GR(r_i, r_j)}{2}, & \text{if } S_{ij}^r \neq 0 \\ GR\left(r_i, r_j\right), & \text{otherwise} \end{cases} \tag{4}$$

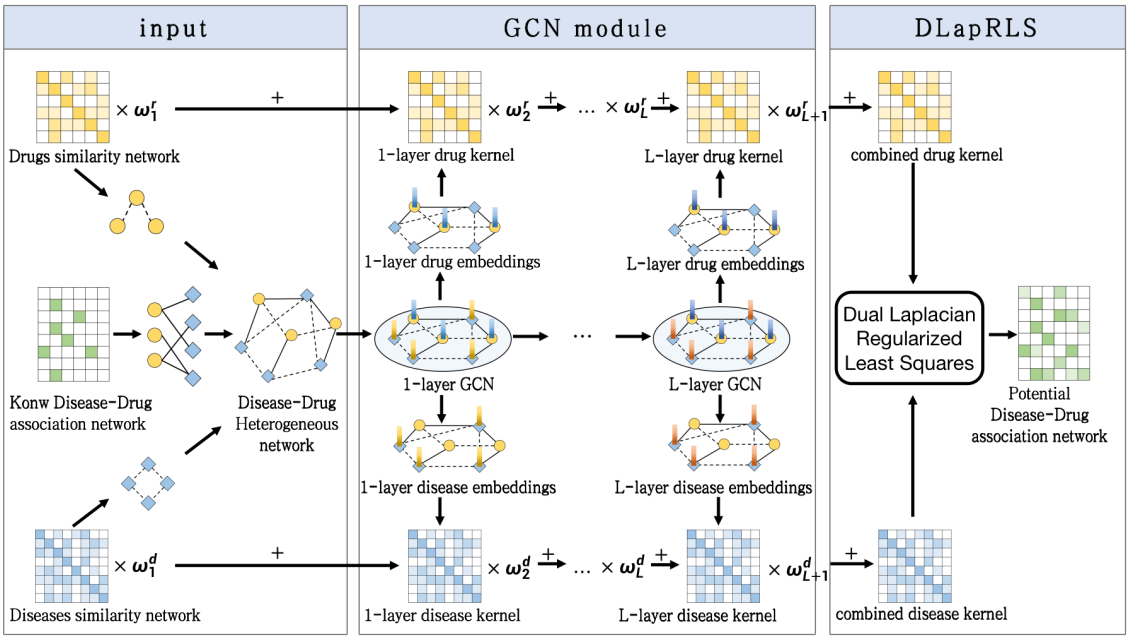

Figure 1: Algorithm flowchart

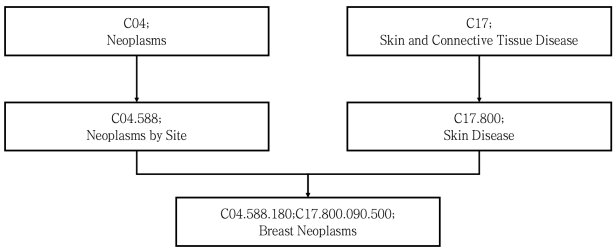

Figure 2: Breast tumor DAG graph structure

Similarly, the disease similarity kernel is constructed by combining the semantic similarity kernel Gaussian kernel similarity of the disease.

Each disease can be represented by a Directed Acyclic Graph (DAG) through standardized mapping of disease names through the Medical Subject Headings (MeSH). Each disease has one or more addresses in the DAG graph, and the address of each disease child node is jointly represented by the addresses of the parent node and the child node. For example, the DAG graph structure of the disease breast tumor is shown in Figure 2.

According to the method proposed by Cui (Wang et al. 2010), the disease-to-disease similarity calculation is based on the DAG structures of the two diseases. Use $DAG(d) = \{\mathcal{N}(d), \mathcal{E}(d)\}$ to represent the hierarchical relationship of diseases, where $\mathcal{N}(d)$ is the node set containing d and its ancestors, and $\mathcal{E}(d)$ is the set of direct connections from the parent node to its child nodes. Based on this DAG structure, the contribution of node n in $DAG(d)$ to the semantic value of disease d is

$$C_d(n) = \begin{cases} 1 & \text{if } n = d \\ \max\{\Delta * C_d(n') \mid n' \in \text{ children of } n\} & \text{if } n \neq d \end{cases}, \tag{5}$$

where Δ is a contribution factor between 0 and 1, which is set to 0.5 here. Define the semantic value of the disease as $DV(d) = \sum_{n \in \mathcal{N}(d)} C_d(n)$. According to the assumption that the more common ancestors, the higher the semantic similarity, use equation 6 to calculate the semantic similarity between diseases d_i and d_j:

$$S_{ij}^d = \frac{\sum_{n \in \mathcal{N}(d_i) \sim \mathcal{N}(d_j)} \left(\mathrm{C}_{d_i}(n) + \mathrm{C}_{d_j}(n)\right)}{\mathrm{DV}(d_i) + \mathrm{DV}(d_j)} \tag{6}$$

The Gaussian interaction profile kernel similarity between diseases is then computed:

$$GD(d_i, d_j) = \exp\left(-\eta_d \left\| Y_{d_i} - Y_{d_j} \right\|^2\right), \tag{7}$$

where y_{d_i} represents the i-th column in the adjacency matrix, that is, the interaction spectrum of the disease d_i, and η_d represents the normalized kernel bandwidth (Van Laarhoven, Nabuurs, and Marchiori 2011), defined as follows:

$$\eta_d = \frac{\eta_d}{\left(\frac{1}{N_d} \sum_{i=1}^{N_d} \left\| Y_{d_i} \right\|^2\right)}, \tag{8}$$

where η_d is the raw bandwidth.

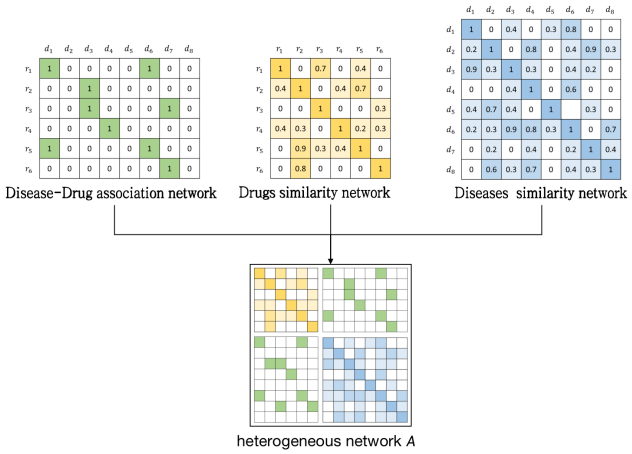

Figure 3: Schematic diagram of adjacency matrix composition

The comprehensive drug similarity $\mathbf{K}_s^d$ of two diseases d_i, d_j is defined as follows:

$$\mathbf{K}_s^d(d_i, d_j) = \begin{cases} \frac{S_{ij}^d + GD(d_i, d_j)}{2}, & \text{if } S_{ij}^d \neq 0 \\ GD(d_i, d_j), & \text{otherwise} \end{cases} \quad (9)$$

In order to integrate network information, a heterogeneous network H including drug-disease association network $Y \in \mathbf{R}^{N_r \times N_d}$, drug similarity network $\mathbf{K}_s^r$, and disease similarity network $\mathbf{K}_s^d$ is constructed, represented by an adjacency matrix A.

$$A = \begin{bmatrix} \mathbf{K}_s^r & Y \\ Y^T & \mathbf{K}_s^d \end{bmatrix} \quad (10)$$

The relationship between the adjacency matrix A and the three kinds of incidence matrices is shown in Figure 3.

Combined Kernel on Graph Embedding

A graph convolutional network (GCN) is a multilayer connected neural network architecture for learning low-dimensional representations of nodes from graph structures.Each layer of the GCN aggregates neighbor information through direct links of the graph, and the reconstructed embeddings are used as inputs for the next layer. For the adjacency matrix A of the heterogeneous network H defined above, the corresponding GCN can be defined as:

$$\mathbf{H}^{(l)} = f\left(H^{(l-1)}, A\right) = \sigma\left(D^{-\frac{1}{2}} A D^{\frac{1}{2}} H^{(l-1)} W^{(l-1)}\right) \quad (11)$$

where $H^{(l)}$ is the l-layer embedding of nodes, where $l = 1, ..., L$;$D = diag\left(\sum_{j=1}^{N_r+N_d} A_{i,j}\right)$ is the diagonal node degree matrix of A; $W^{(l)} \in \mathbf{R}^{(\mathbf{N_d}+\mathbf{N_m}) \times \mathbf{k_1}}$ is a weight matrix for the l-th neural network layer and k_l is the dimensionality of embeddings of l-th layer GCN; $\sigma(\cdot)$ is a non-linear activation function.In this paper, $\mathbf{ReLU}$ (modified linear unit) is used as the activation function. For the first layer of GCN, the initial embedding $H^{(0)}$ is constructed as follows:

$$\mathbf{H^{(0)}} = \begin{bmatrix} 0 & Y \\ Y^T & 0 \end{bmatrix} \quad (12)$$

The multi-layer GCN model captures different structural information of heterogeneous networks and can compute multiple embeddings. For example, the first layer captures direct link information, and higher layers capture multi-hop neighbor information (higher-order proximity) by iteratively updating the embeddings (Van Laarhoven, Nabuurs, and Marchiori 2011).Since the embeddings at each layer represent different information, in this paper, the embeddings at each layer are used as different feature vectors to compute multiple kernel matrices.

For the embedding of each layer $H_i(i = 1, ..., L)$, we can divide it to two parts, the first N_r lines are used as drug embeddings H_i^r, and the last N_d lines are used as microbe embeddings H_i^d. The kernel matrix for each layer of drug and disease embedding was calculated using the Gaussian interaction profile (GIP) as follows:

$$K_{h_l}^r = \exp\left(-\gamma_{h_l} ||H_l^r(i) - H_l^r(j)||^2\right) \quad (13)$$

$$K_{h_l}^d = \exp\left(-\gamma_{h_l} ||H_l^d(i) - H_l^d(j)||^2\right) \quad (14)$$

where $H_i^r(i)$ and $H_i^d(i)$ are profiles of the i-row in the l-layer drug and microbe embeddings; r_{h_l} denotes the corresponding bandwidth.

Since different embeddings represent various structural information, the kernel composed of different embeddings will represent the similarity between nodes from different views. Combined with existing similarity matrices, we have the kernel sets of drug feature space $S^r = \{K_s^r, K_{h_1}^r, ..., K_{h_L}^r\}$ and microbe feature space $S^d = \{K_s^d, K_{h_1}^d, ..., K_{h_L}^d\}$.

To improve the prediction performance, the above kernel matrices are combined in two spaces separately by a weighted sum method. The combined kernels are defined as follows:

$$K_r = \sum_{i=1}^{L+1} \omega_i^r S_i^r \quad (15)$$

$$K_d = \sum_{i=1}^{L+1} \omega_i^d S_i^d \quad (16)$$

where S_i^r and S_i^d are ith kernels in drug and microbe kernel set, ω_i^r and ω_i^d are the corresponding weight of each kernel. Here, we set $\omega_i^r = \omega_i^d = \frac{1}{l+1}$.

Classification Prediction Algorithm

In this paper, we apply the dyadic Laplace regularized least squares (DLapRLS) (Ding, Tang, and Guo 2020) framework to predict associations. DLapRLS is a model based on two eigenspace kernel matrices and aims to find the optimal prediction matrix F^* by minimizing the following objective function:

$$minJ = ||K_r \alpha_r + (K_d \alpha_d)^T - 2Y_{train}||_F^2 \quad (17)$$

where $||\cdot||_F^2$ is the Frobenius norm, $Y_{train} \in R^{N_r \times N_d}$ is the adjacency matrix for microbe-drug associations in the training set; $\alpha_r \in R^{N_r \times N_d}$ and $\alpha_d^T \in R^{N_r \times N_d}$ are trainable matrices; $K_r \in R^{N_r \times N_r}$ and $K_d \in R^{N_d \times N_d}$ are fused kernels in two feature spaces respectively.

To minimize the distance between the potential feature vectors of two adjacent drugs or adjacent diseases, add graph regularization for drugs and diseases, respectively:

$$\min_{\alpha_r} \sum_{i,j}^{n} K_r(i,j) ||\alpha_r^i - \alpha_r^j||^2 = tr\left(\alpha_r^T L_r \alpha_r\right) \quad (18)$$

$$\min_{\alpha_d} \sum_{p,q}^{n} K_d(p,q) ||\alpha_d^p - \alpha_d^q||^2 = tr\left(\alpha_d^T L_d \alpha_d\right) \quad (19)$$

where $i,j = 1,2,...N_r$, $p,q = 1,2,...,N_d$; α_r^i is the i-th row vector of α_r and α_d^p is the p-th row vector of α_r; $L_r \in R^{N_r \times N_r}$ and $L_d \in R^{N_d \times N_d}$ are the normalized Laplacian matrices, respectively:

$$L_r = D_r^{-\frac{1}{2}} \Delta_r D_r^{\frac{1}{2}}, \Delta_r = D_r - K_r \quad (20)$$

$$L_d = D_d^{-\frac{1}{2}} \Delta_d D_d^{\frac{1}{2}}, \Delta_d = D_d - K_d \quad (21)$$

where $D_r = diag\left(\sum_{j=1}^{N_r} K_r(i,j)\right)$ and $D_d = diag\left(\sum_{j=1}^{N_d} K_d(i,j)\right)$.

The final objective function is:

$$\begin{aligned} minJ = &||K_r\alpha_r + (K_d\alpha_d)^T - 2Y_{train}||_F^2 \\ &+\lambda_r tr\left(\alpha_r^T L_r \alpha_r\right) + \lambda_d tr\left(\alpha_d^T L_d \alpha_d\right) \end{aligned} \quad (22)$$

The prediction matrix F^* of the drug-disease association from the two feature spaces is as follows:

$$F^* = \frac{K_r\alpha_r + (K_d\alpha_d)^T}{2} \quad (23)$$

Optimization

There are two types of parameters in the model that need to be optimized. The first category is the GCN parameters, which are optimized by the Adam optimizer (Kingma and Ba 2014) to minimize the loss function; the second category is the parameters of DLapRLS, where the iteration function is obtained directly by calculating the partial derivatives.

In optimizing α_r, first assume that α_r is known, and then calculate the partial derivative of the objective function with respect to α_r:

$$\frac{\partial J}{\alpha_r} = 2K_r\left(K_r\alpha_r + \alpha_d^T K_d^T - 2Y_{train}\right) + 2\lambda_r L_r \alpha_r \quad (24)$$

By letting $\frac{\partial J}{\alpha_r} = 0$,we can obtain:

$$\alpha_r = \left(K_r\alpha_r + \lambda_r L_r\right)^{-1} K_r[2Y_{train} - \alpha_d^T K_d^T] \quad (25)$$

Similarly, we calculate the partial derivative of the loss function with respect to α_d as follows:

$$\frac{\partial J}{\alpha_d} = 2K_d\left(K_d\alpha_d + \alpha_r^T K_r^T - 2Y_{train}\right) + 2\lambda_d L_d \alpha_d \quad (26)$$

By letting $\frac{\partial J}{\alpha_d} = 0$,we can obtain:

$$\alpha_d = \left(K_d K_d + \lambda_d L_d\right)^{-1} K_d[2Y_{train} - \alpha_m^T K_m^T] \quad (27)$$

The inputs to the model are the known association matrix Y and the similarity matrices K_s^r and K_d^r of drugs and diseases. In training, firstly, all trainable parameters are randomly initialized, and the heterogeneous network is constructed according to the input matrix to initialize the embedding $H^{(0)}$. In each iteration, the GCN computes the embedding of each layer of the node by forward propagation and calculates the kernel matrix based on the embedding; then multiple kernel matrices are fused in each of the two spaces; then the embedding of the GCN is updated by back propagation and the parameters of DLapRLS are updated by iterative functions. The prediction matrix F^* is output after iterative update.

Results

Experimental setting

The k-fold cross validation is widely used to evaluate the predictive performance of a model. In the cross-validation process, the correlated dataset is first divided into k parts randomly and equally, and one of the parts is selected as the test set each time, and the other k-1 parts are used as the training set to train and test the model for a total of k validations. In this paper, a 5-fold cross-test is used and we use two evaluation indicators: the area under thereceiver operating characteristic curve (AUC) and the area underthe precision recall curve (AUPR).

Parameters evaluation

In the study of this paper, the main parameters that affect the prediction effect of the model are λ_r, λ_d and γ_{hl} (l=1,2,...,L). The parameter γ_{hl} controls different K_{hl}^r and K_{hl}^d, and fixes λ_r and λ_d, when γ_{hl} is in 2^{-3},..., 2^0,..., 2^3 changes, AUPR changes accordingly. Figure 4, Figure 5, and Figure 6 respectively show the impact of changes in γ_{hl} (l=1,2,3) on the prediction performance. It can be observed that the prediction performance of the model is poor when γ_{hl} is large. For γ_{h1}, AUPR decreases with the increase of γ_{h1}, and reaches the optimal value at the minimum value of γ_{h1}. When γ_{h2} and γ_{h3} are small, there is little change, but overall AUPR also decreases with the increase of its value. In order to obtain the best prediction effect, this paper selects $\gamma_{h1} = 2^{-3}$, $\gamma_{h2} = 2^{-2}$, $\gamma_{h3} = 2^{-3}$ as model parameters.

λ_r and λ_d represent the weights of graph regularization items in DLapRLS, with the highest AUPR as the parameter selection index, as shown in Table 2, when $\lambda_r = 2^{-2}$, $\lambda_d = 2^{-1}$, AUPR reaches a maximum value of 0.5379.

Evaluation of the role of each part of the model

The MKGCN model is based on graph convolutional networks, where different layers of GCNs can yield different node embeddings, and then multiple kernel matrices based on different embedding information. This paper discusses experimentally the effect of kernel matrices generated by different layers and known similarity matrices, as well as the

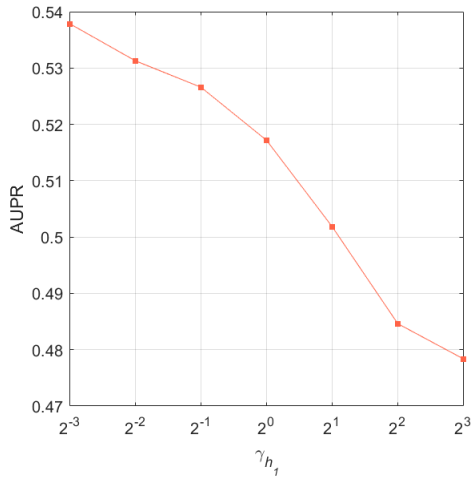

Figure 4: AUPR of the model under different γ_{h1}

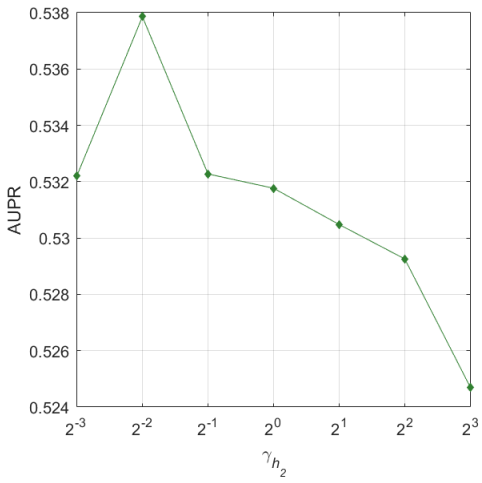

Figure 5: AUPR of the model under different γ_{h2}

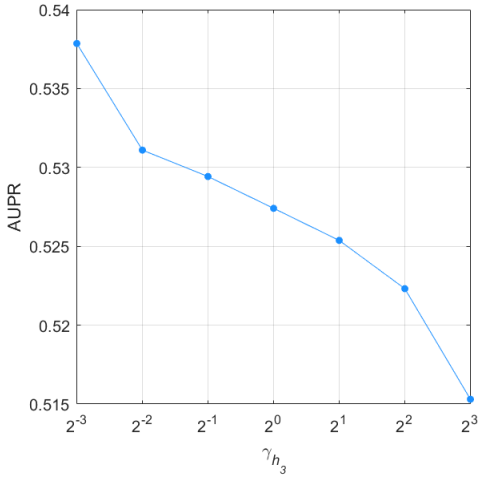

Figure 6: AUPR of the model under different γ_{h3}

λ_r	λ_d			
	2^{-3}	2^{-2}	2^{-1}	2^{0}
2^{-3}	0.5167	0.5257	0.5291	0.5274
2^{-2}	0.5201	0.5201	**0.5379**	0.5375
2^{-1}	0.5180	0.5312	0.5373	0.5368
2^{0}	0.5144	0.5277	0.5333	0.5318

Table 2: AUPR values at different λ_r and λ_d

difference between single and multiple kernels for modalities.

The model uses a three-layer GCN, and research (Yu et al. 2021) shows that the performance of the GCN encoder decreases as the number of convolution layers increases. hl+MKGCN denotes the prediction using the kernel matrix computed by the MKGCN model on the lth layer (l=1,2,3) of embeddings. In addition, to verify the effectiveness of the graph convolution in extracting features, the prediction is performed directly using the original similarity matrices K_s^r and K_s^d using the s+MKGCN representation. mean+MKGCN represents the fusion of the above four kernel matrices by assigning average weights to them.

Models	AUPR	AUC
h1+MKGCN	0.5127	0.8548
h2+MKGCN	0.4659	0.8355
h3+MKGCN	0.4119	0.8079
s+MKGCN	0.3954	0.7875
mean+MKGCN	**0.5379**	**0.8609**

Table 3: Prediction effects of models based on different kernels

The AUCs and AUPRs of all models under 5-fold cross-check are shown in Table 3. In the single kernel models, the AUC and AUPR of h1+MKGCN and h2+MKGCN are higher than those of h3+MKGCN and s+MKGCN, which means that the kernel matrices generated by the first and second layers of the GCN contain more information than the third layer and the known similarity matrix and therefore achieve better results. It can be seen that the kernel matrix calculated based on the GCN embedding is an effective way to describe the relationships between the nodes and can provide additional information to the model to improve the prediction. On the other hand, comparing the MKGCN with the single kernel model, the MKGCN outperforms the single kernel model, indicating that the multi-core model can combine more information for prediction. Therefore, in the next experiments, the average weighting method is chosen to construct the MKGCN model in this paper.

Comparative experiments

To further evaluate the predictive performance of the MKGCN model, this paper compares MKGCN with two more advanced drug-disease association prediction methods. The two methods are SCMFDD (Zhang et al. 2018b) and LAGCN (Yu et al. 2021) respectively. As shown in Tables 4, the three methods have similar results in terms of AUC metrics, but in terms of AUPR metrics, MKGCN has a significant advantage of up to 0.5379, compared to 0.2399 and 0.3008 for SCMFDD and LAGCN respectively, with MKGCN improving 124.2% and 78.8% relative to the two methods respectively. It can therefore be seen that the prediction performance of the GCN-based method outperforms that of the matrix decomposition-based method, indicating that the GCN can aggregate feature information of the network topology well. And also using multiple convolutional layers of GCN to embed information, the multi-core learning approach works better than the attention mechanism, indicating that the multiple sources of information provided by the kernel matrix can significantly improve the prediction ability of the model.

Models	AUPR	AUC
SCMFDD	0.2399	0.8619
LAGCN	0.3008	0.8731
MKGCN	**0.5379**	**0.8609**

Table 4: Prediction results under 5-fold cross-validation of different methodological models

To illustrate the fitness of MKGCN on different datasets, a 5-fold cross-validation was performed on three additional datasets (Fdataset (Gottlieb et al. 2011), Cdataset (Luo et al. 2016) and DNdataset (Martinez et al. 2015)) in this paper. Among them, Fdataset includes 593 drugs from the DrugBank database (Wishart et al. 2018), 313 diseases and 1933 known associations from the OMIM dataset (Hamosh et al. 2002); Cdataset contains 663 drugs collected from DrugBank, 409 diseases from the OMIM database and 2352 known drug-disease associations; DNdataset contains 1490 drugs from DrugBank, 4516 diseases and 1008 known drug-disease associations from Disease Ontology (Schriml et al. 2012). Information on the dataset is shown in Tables 5.

Datasets	Drugs	Diseasess	Associations	Sparsity
Our dataset	269	598	18416	0.11448
Fdataset	593	313	1993	0.01041
Cdataset	663	409	2352	0.00867
DNdataset	1490	4516	1008	0.00015

Table 5: Statistical information on the datasets

On Fdataset, MKGCN (AUPR: 0.5234, AUC: 0.9001) achieved the best results, with a significant improvement over SCMFDD (AUPR: 0.0259, AUC: 0.7745) and LAGCN (AUPR: 0.0830, AUC: 0.7689); on Cdataset, the MKGCN model with AUPR and AUC of 0.6307 and 0.9160, respectively, predicted better than SCMFDD (AUPR: 0.0235, AUC: 0.7903) and LAGCN (AUPR: 0.0957, AUC: 0.7615); on DNdataset, the AUPR and AUC of the MKGCN model were were 0.3662 and 0.8831, with AUPRs 0.2562 and 0.1106 higher than those of SCMFDD (AUPR: 0.1100, AUC: 0.9413) and LAGCN (AUPR: 0.2556, AUC: 0.7468), respectively. The relevant results are shown in Table 6.

Datasets	Models	AUPR	AUC
Fdataset	SCMFDD	0.0259	0.7745
	LAGCN	0.0830	0.7689
	MKGCN	**0.5234**	**0.9001**
Cdataset	SCMFDD	0.0235	0.7903
	LAGCN	0.0957	0.7615
	MKGCN	**0.6307**	**0.9160**
DNdataset	SCMFDD	0.1100	**0.9413**
	LAGCN	0.2556	0.7468
	MKGCN	**0.3662**	0.8831

Table 6: Prediction results of different methodological models on various datasets

The above experiments illustrate that the MKGCN model has good performance in drug-disease association prediction and can be generalised to different datasets.

Case Study

To illustrate the predictive effect of the model biologically, the MKGCN model was constructed using the entire association dataset, and then the known drug-disease associations in the prediction matrix were removed and the 10 drug-disease associations with the highest prediction scores were selected. The prediction results are shown in Tables 7, of which six associations have biomedical support in the literature. For example, morphine, currently used clinically primarily for analgesia and cough suppression, was predicted to treat precursor cell lymphocytic leukaemia, while the antipsychotic drug risperidone, was predicted to treat angioedema, and the anti-folate antineoplastic agent methotrexate, was predicted to treat liver failure.

Associated diseases for carbamazepine and associated drugs for breast tumours were then examined, with the results shown in Tables 8, where the percentage of associations with literature support was 80% and 100% respectively. Carbamazepine is an anticonvulsant drug used primarily for the treatment of epilepsy and neuropathic pain, and predictions found associations with coagulation disorders, squamous cell carcinoma, and nephrogenic uremic syndrome as well. Breast cancer is a common malignancy in women, and MKGCN predicts an association with drugs such as adriamycin and vitamin A acid.

Discussion and conclusion

Determining drug-disease associations provides important information for new drug development, however, the process

Drug	Disease	Evidence
Morphine	Leukaemia-Lymphoma	Y
Promethazine	Coagulation disorders	N
Risperidone	Angioedema	Y
Methotrexate	Liver Failure	Y
Phenytoin	Solitary neuropathy	N
Colotine	Neurodegeneration	Y
Quetiapine fumarate	Leukaemia-Lymphoma	Y
Sertraline	Neurodegeneration	Y
Valproic acid	Hyperhidrosis	N
Sertraline	Nephrogenic Enuresis	N

Table 7: Top 10 drug-disease associations predicted by MKGCN

of determining drug-disease associations by wet-lab methods is very time-consuming and expensive. Currently, many drug-disease relationships are still unknown and developing a new method to predict drug-disease associations that have not yet been discovered is a very important and urgent topic. In this paper, a multicore fusion model (called MKGCN) based on graph convolutional networks is developed for predicting drug-disease associations.

The advantage of multicore fusion is to improve the predictive power of the prediction model by combining multicore information to fully describe the relationship between samples. However, in traditional multicore learning algorithms, the kernel matrix is constructed by extracting features from multiple known sources of information in the sample itself, and this approach is not suitable for samples with fewer sources of information. Therefore, unlike traditional multicore learning, this paper uses GCN to construct multiple kernel matrices based on the embedding of different layers of networks to solve the above problem and achieve the purpose of making full use of multiple information.

The experimental results of cross-validation show that MKGCN is an effective prediction method, especially when compared with existing models, MKGCN achieves the best AUPR values on multiple data sets. In addition, the case study also confirmed the reliability of the predicted potential drug-disease associations in terms of bio-interpretation.

In summary, the present model can be used as an effective method for predicting drug-disease associations, providing new ideas for drug repositioning calculations, and providing computational aids for clinical trials.

In future work, in-depth research and improvement can be conducted in the following aspects.

- **Selection of plausible negative samples:** there is a lack of drug-disease negative pairs in current publicly available databases and published literature, and drug-disease pairs that are not associated by default in prediction experiments may also have associations that are not experimentally validated, thus affecting the prediction effectiveness of the model. Further exploration of effective methods to identify negative samples in combination with more pathology knowledge and clinical information is needed.

Drug	Disease	Evidence
Carbamazepine	Coagulation disorders	Y
	Squamous cell carcinoma	Y
	Nephrogenic Urogenital Disease	Y
	Hot flushes	N
	Hyperammonia	Y
	Hyperthyroidism	Y
	Myopathy	Y
	Nephrosclerosis	Y
	Neurodegeneration	Y
	Gastrointestinal signs and symptoms	N
Disease	**Drug**	**Evidence**
Breast neoplasms	Adriamycin	Y
	Vitamin A acid	Y
	Algocytidine	Y
	Vincristine	Y
	Mitoxantrone	Y
	Sorafenib	Y
	Dexamethasone	Y
	Tamoxifen	Y
	Prednisone	Y
	Quercetin	Y

Table 8: MKGCN predicts the top 10 related diseases (related drugs) with the highest score for a drug (disease)

- **Integration of feature information from multiple data sources:** features such as chemical structure, side effect information, and interactions of drugs can be used to construct similarity networks. Similarly, more information on disease features can be extracted based on the association of diseases with microorganisms, genetic information of diseases, etc. Therefore, adding more feature information can be considered afterwards to construct more accurate and better similarity networks of diseases and drugs.
- **Optimization of multiple kernel fusion parameters:** The ablation experiment found that each kernel matrix has different effects on the prediction effect. In this paper, considering the simplicity of model calculation, the average weighting method is directly used to fuse each kernel matrix. The idea of attention mechanism can be borrowed to change the weight coefficients in multicore fusion into adaptive parameters, so as to improve the prediction effect of the model.

References

Aliper, A.; Plis, S.; Artemov, A.; Ulloa, A.; Mamoshina, P.; and Zhavoronkov, A. 2016. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7): 2524–2530.

Booth, B.; and Zemmel, R. 2004. Opinion/Outlook: Prospects for productivity. *Nature reviews. Drug discovery*, 3: 451–6.

Chen, B.; Ding, Y.; and Wild, D. J. 2012. Assessing drug target association using semantic linked data. *PLoS computational biology*, 8(7): e1002574.

Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; and Mattingly, C. J. 2017. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1): D972–D978.

Ding, Y.; Tang, J.; and Guo, F. 2020. Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. *Knowledge-Based Systems*, 204: 106254.

Gottlieb, A.; Stein, G. Y.; Ruppin, E.; and Sharan, R. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1): 496.

Hamosh, A.; Scott, A. F.; Amberger, J.; Bocchini, C.; Valle, D.; and McKusick, V. A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 30(1): 52–55.

Hu, S.; Zhang, C.; Chen, P.; Gu, P.; Zhang, J.; and Wang, B. 2019. Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC bioinformatics*, 20(25): 1–12.

Huang, Y.-a.; Hu, P.; Chan, K. C.; and You, Z.-H. 2020. Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics*, 36(3): 851–858.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, J.; Zhang, S.; Liu, T.; Ning, C.; Zhang, Z.; and Zhou, W. 2020. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics*, 36(8): 2538–2546.

Luo, H.; Wang, J.; Li, M.; Luo, J.; Peng, X.; Wu, F.-X.; and Pan, Y. 2016. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*, 32(17): 2664–2671.

Martinez, V.; Navarro, C.; Cano, C.; Fajardo, W.; and Blanco, A. 2015. DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artificial intelligence in medicine*, 63(1): 41–49.

Plenge, R. M.; Scolnick, E. M.; and Altshuler, D. 2013. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8): 581–594.

Scannell, J.; Blanckley, A.; Boldon, H.; and Warrington, B. 2012. Diagnosing the Decline in Pharmaceutical RD Efficiency. *Nature reviews. Drug discovery*, 11: 191–200.

Schriml, L. M.; Arze, C.; Nadendla, S.; Chang, Y.-W. W.; Mazaitis, M.; Felix, V.; Feng, G.; and Kibbe, W. A. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1): D940–D946.

Van Laarhoven, T.; Nabuurs, S. B.; and Marchiori, E. 2011. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21): 3036–3043.

Wang, D.; Wang, J.; Lu, M.; Song, F.; and Cui, Q. 2010. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, 26(13): 1644–1650.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1): D1074–D1082.

Wu, C.; Gudivada, R. C.; Aronow, B. J.; and Jegga, A. G. 2013. Computational drug repositioning through heterogeneous network clustering. *BMC systems biology*, 7(5): 1–9.

Yang, K.; Zhao, X.; Waxman, D.; and Zhao, X.-M. 2019. Predicting drug-disease associations with heterogeneous network embedding. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12): 123109.

Yu, Z.; Huang, F.; Zhao, X.; Xiao, W.; and Zhang, W. 2021. Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics*, 22(4): bbaa243.

Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; and Cheng, F. 2019. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24): 5191–5198.

Zhang, W.; Yue, X.; Lin, W.; Wu, W.; Liu, R.; Huang, F.; and Liu, F. 2018a. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics*, 19(1): 1–12.

Zhang, W.; Yue, X.; Lin, W.; Wu, W.; Liu, R.; Huang, F.; and Liu, F. 2018b. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics*, 19(1): 1–12.

Zitnik, M.; Agrawal, M.; and Leskovec, J. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13): i457–i466.