Graph Convolutional Autoencoder and Generative Adversarial Network-Based Method for Predicting Drug-Target Interactions

Chang Sun, Ping Xuan, Tiangang Zhang, and Yilin Ye

Abstract—The computational prediction of novel drug-target interactions (DTIs) may effectively speed up the process of drug repositioning and reduce its costs. Most previous methods integrated multiple kinds of connections about drugs and targets by constructing shallow prediction models. These methods failed to deeply learn the low-dimension feature vectors for drugs and targets and ignored the distribution of these feature vectors. We proposed a graph convolutional autoencoder and generative adversarial network (GAN)-based method, GANDTI, to predict DTIs. We constructed a drug-target heterogeneous network to integrate various connections related to drugs and targets, i.e., the similarities and interactions between drugs or between targets and the interactions between drugs and targets. A graph convolutional autoencoder was established to learn the network embeddings of the drug and target nodes in a low-dimensional feature space, and the autoencoder deeply integrated different kinds of connections within the network. A GAN was introduced to regularize the feature vectors of nodes into a Gaussian distribution. Severe class imbalance exists between known and unknown DTIs. Thus, we constructed a classifier based on an ensemble learning model, LightGBM, to estimate the interaction propensities of drugs and targets. This classifier completely exploited all unknown DTIs and counteracted the negative effect of class imbalance. The experimental results indicated that GANDTI outperforms several state-of-the-art methods for DTI prediction. Additionally, case studies of five drugs demonstrated the ability of GANDTI to discover the potential targets for drugs.

Index Terms—Adversarial regularization, drug-target interaction, generative adversarial network, graph convolutional autoencoder, LightGBM

1 Introduction

Drugs exert their efficacy by interacting with various molecular targets via drug-target interaction (DTI). Proteins are one important group of such molecular targets [1]. Drugs affect disease conditions by enhancing or inhibiting expression of the target proteins to which they bind [2], [3]. Previous studies have demonstrated that the drugs approved by the Food and Drug Administration (FDA) can interact with several targets [4]. Existing drugs with potentially unobserved targets likely have unknown indications [5]. However, determining DTIs through biological experiments is time consuming, laborious and costly [6]. Many studies have therefore begun predicting novel DTIs via computational methods [7], [8], [9]. These studies can provide biologists with DTI candidates to reduce the workload of the wet-lab experiments.

Traditional methods for DTI prediction can be categorized into docking-based methods and ligand-based approaches [10]. Docking-based methods require the 3D structures of the target proteins. As the structural information is not known for all targets, the performance of these methods was limited [11], [12]. Ligand-based methods compare the protein with the unknown ligand with a set of proteins with known ligands [13]. These approaches do not perform well when the number of known ligands is insufficient.

In recent years, much research has begun predicting DTIs from a network perspective [14], [15]. This kind of method analyzes potential DTIs by integrating various information in the heterogeneous drug-target network. Chen *et al.* applied random walk on a drug-target heterogeneous network to predict DTIs [16]. Ezzat *et al.* developed a graph regularization-based matrix factorization model to predict potential DTIs [17]. This improved the density of the drug-target interaction matrix by inferring possible DTIs, making the prediction result more accurate. Luo *et al.* extracted effective information from the adjacency matrices of the drug and target networks with a singular value decomposition algorithm [18]. This matrix factorization-based method was named DTINet. However, both random walk and matrix factorization are shallow models and therefore cannot fully explore the deep relationships between drugs and their targets.

Bleakley and Yamanishi constructed a bipartite local model and forecasted DTIs with a support vector machine (SVM) [19]. Lee and Nam proposed a DTI prediction method based on restart random walk (RWR) and k-Nearest-Neighbors (KNN) [20]. They applied RWR to the drug and target networks. The drug and target features were given different weights based on the RWR results. A KNN model was used to calculate the interaction score for each drug-target pair. The traditional machine learning model such as SVM, KNN

- C. Sun, P. Xuan, and Y. L. Ye are with the Department of Computer Science and Technology, Heilongjiang University, Harbin, Heilongjiang 150080, China. E-mail: {sunchangcn, yeyilincn}@outlook.com, xuanping@hlju.edu.cn.
- T. G. Zhang is with the Department of Mathematical Science, Heilongjiang University, Harbin, Heilongjiang 150080, China. E-mail: zhang@hlju.edu.cn.

Manuscript received 7 Sept. 2019; revised 24 May 2020; accepted 28 May 2020. Date of publication 1 June 2020; date of current version 3 Feb. 2022.
(Corresponding author: Ping Xuan.)
Digital Object Identifier no. 10.1109/TCBB.2020.2999084

usually only uses the same quantity of unknown DTIs as known DTIs to train the model. Nevertheless, there is a serious class imbalance between the two. The performance of such models was therefore limited, as most unknown DTIs were abandoned.

To predict novel DTIs, Xuan *et al.* developed an ensemble learning method called DTIGBDT that can train the model with all samples in the dataset [21]. It extracted path category-based feature vectors to incorporate the topological information of the drug-target heterogeneous network. A gradient boosting decision tree-based model was established to analyze drug-target associations. Meanwhile, DTIGBDT did not deeply learn the low-dimensional feature vectors for drugs and targets.

In this study, we developed a new method named GANDTI to accurately predict DTIs. GANDTI deeply integrated the topological information and node attributes of the drug-target heterogeneous network through a graph convolutional autoencoder [22]. The embedded representations of the drug and target nodes in the heterogeneous network were obtained by the encoder. In addition, the embedded representations were altered to match a Gaussian distribution by a generative adversarial network [23], which can improve the robustness of the encoder [24]. The DTI propensities were calculated using a LightGBM-based classifier [25]. As an ensemble learning [26] model based on decision tree [27], LightGBM can completely utilize unknown DTIs and efficiently release eradicate the negative effects of class imbalance.

2 Materials and Method

Our primary aim was to predict possible DTIs by analyzing drug and target attributes, as well as investigating the interactions between the two. We therefore constructed a drug-target heterogeneous network and extracted the edge information (network topology) and node information (node attributes) of the network. An adversarial graph convolutional encoder was used to learn the feature representation of each node in the network. The interaction scores between drug-target pairs were calculated via a LightGBM-based classifier.

2.1 Dataset

The dataset for DTI prediction was obtained from a previous study involving 549 drugs and 424 targets [18]. This dataset includes five types of data: (1) chemical structure information of 549 drugs; (2) 10,036 drug-drug interactions (DDIs); (3) primary sequences of 424 target proteins; (4) 7,363 target-target interactions (TTIs); and (5) 1,923 known DTIs. The protein sequences were extracted from the Human Protein Reference Database (HPRD) [28]. The remaining data were obtained from the DrugBank database [29].

2.2 Construction of Drug-Target Heterogeneous Network

We constructed the DDI network drugNet and the TTI network targetNet. $D = \{d_1, d_2, \ldots, d_m\}$ was used to represent m drug nodes in drugNet. Each edge in drugNet indicated a known interaction between the two drug nodes connected by this edge. Similarly, we used $T = \{t_1, t_2, \ldots, t_n\}$ to represent n target nodes in targetNet. The edge was added when the two target nodes had a known interaction. In addition, if there was a known interaction between a drug node and a target node, an edge was added between the two nodes. Thus, based on the three types of interactions in the dataset (DDIs, TTIs, and DTIs), we constructed the drug-target heterogeneous network dtNet.

The topological information of drugNet, targetNet, and dtNet can be represented by the adjacency matrices of these networks. For example, we used $A^D \in R^{m\times m}$ to denote the adjacency matrix of drugNet. $A^D_{i,j} = 1$ when there was an edge between the drug d_i and d_j, otherwise $A^D_{i,j} = 0$. Similarly, the adjacency matrices of targetNet and dtNet were represented by $A^T \in R^{n\times n}$ and $Y \in R^{m\times n}$, respectively. We calculated the Jaccard similarity coefficients [30] between the drugs based on their chemical structure and constructed the similarity matrix for drugs, which were denoted by $S^D \in R^{m\times m}$. The primary sequences of the targets were used to calculate the Smith-Waterman score [31] between the targets, and to construct a similarity matrix $S^T \in R^{n\times n}$. The values of the elements in S^D and S^T were scaled into [0, 1] by row normalization, and were used to describe the similarity between two drugs or targets. It is generally believed that the closer $S^D(i,j)$ (or $S^T(i,j)$) is to 1, the more similar d_i and d_j (or t_i and t_j).

2.3 Adversarial Graph Convolutional Autoencoder

We aimed to determine the low-dimensional embedding representations for all drugs and targets, such as the drug and target feature vectors, in dtNet. The feature vectors were fed into the classifier for calculating the interaction scores between the drugs and targets.

The topological information matrix of dtNet consists of a drug-drug interaction matrix A^D, a target-target interaction matrix A^T, and a drug-target interaction matrix Y. As seen in Fig 1, we connected the three matrices and obtained the topological relationship matrix $\tilde{A} \epsilon R^{(m+n)\times(m+n)}$ of dtNet. The self-attribute $\tilde{X}_i$ of the drug node d_i was obtained by concatenating S^D_i and Y_i, where S^D_i is the similarity vector between d_i and other drugs in drugNet, and Y_i is the interaction vector between d_i and other targets in dtNet. Similarly, the self-attribute $\tilde{X}_j$ of the target t_j was obtained by connecting the interaction vector Y_i and the similarity vector S^T_j.

2.3.1 Graph Convolution Encoder

To deeply integrate the topological information of dtNet and the attributes of drug nodes and target nodes, we designed a graph convolutional network (GCN)-based encoder for our network (Fig 2(a)). This encoder has two GCN layers. In order to exploit the information of each node about itself that in the drug-target heterogeneous network into account, we set $A' = \tilde{A} + I$. The topological information matrix $\tilde{A}$ was normalized by graphing a Laplacian matrix to obtain $\bar{A} \epsilon R^{(m+n)\times(m+n)}$, where m is the number of drugs, and n is the number of targets. $\bar{A}$ can be calculated as follows:

$$\bar{A} = \tilde{D}^{-\frac{1}{2}} A' \tilde{D}^{-\frac{1}{2}}. \tag{1}$$

where $\tilde{D}_{ii} = \sum_j A'_{ij}$, and I is the identity matrix. If we want to project the feature vectors of each drug and target node in dtNet into k-dimensional, then the embedding representation matrix $Z \in R^{(m+n)\times k}$ can be calculated as follows:

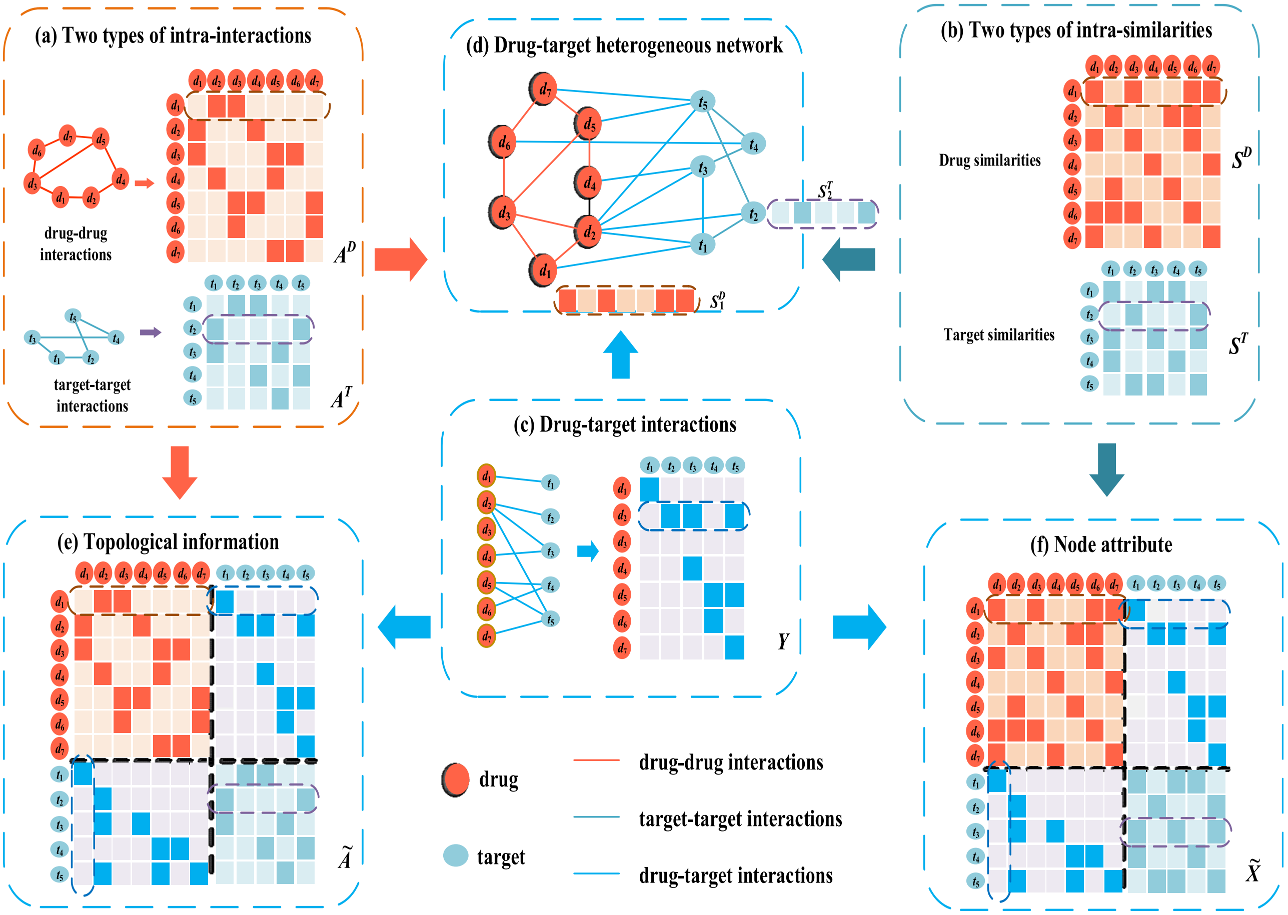

(a) Two types of intra-interactions
drug-drug interactions
A^D
target-target interactions
A^T
(d) Drug-target heterogeneous network
S_2^T
S_1^D
(b) Two types of intra-similarities
Drug similarities
S^D
Target similarities
S^T
(c) Drug-target interactions
Y
(e) Topological information
$\tilde{A}$
(f) Node attribute
$\tilde{X}$
drug
target
drug-drug interactions
target-target interactions
drug-target interactions
d_1 d_2 d_3 d_4 d_5 d_6 d_7
t_1 t_2 t_3 t_4 t_5

Fig. 1. Extraction of the topological information and node attributes from the drug-target heterogeneous network.

$$Z = \phi_1 \left(\bar{A}\ \phi_2(\bar{A}\tilde{X}W_1)W_2\right), \tag{2}$$

where $W_1\epsilon R^{(m+n)\times l}$ and $W_2\epsilon R^{l\times k}$ are the weight matrices of the first and second GCN layers, respectively. l is the dimension of the feature vectors of each drug and target in the feature map, outputted by the first GCN layer. $\phi(\cdot)$ is the activation function, and $\phi_1\ (t) = \mathrm{Relu}\ (t) = \max(0,t)$, $\phi_2\ (t) = sigmoid\ (t) = \frac{1}{1+e^t}$. The range of the elements in Z is [0, 1]. The first m rows of matrix Z represent the low dimensional feature vectors of the m drugs, and the last n rows are the low-dimensional vectors of n targets. Therefore, Z is the low-dimensional representation which deeply fuse the topological information and the node attruibute.

2.3.2 Decoder and Optimizer

The purpose of the decoder was to reconstruct the topological information matrix $\tilde{A}$ of dtNet using the embedding representation matrix Z. The element $\hat{A}(i,j)$ in the reconstructed matrix $\hat{A}\ \epsilon R^{(m+n)\times(m+n)}$ represents the interaction propensity between drug d_i (or target t_i) and another drug d_j (or target t_j), which was predicted by the decoder. This can be calculated by Equation (3):

$$\hat{A}(i,j) \ = \ sigmoid\left(z_i \cdot {z_j}^{\mathrm{T}}\right), \tag{3}$$

where z_i and z_j are low-dimensional feature vectors of node i and node j, respectively, ${z_j}^{\mathrm{T}}$ is the transpose of z_j. The more consistent the feature distributions of the two nodes in the low-dimensional feature space were, the larger the inner product of the vectors corresponding to the two nodes, and the higher the prediction score of the decoder. To make the topological information matrix $\tilde{A}$ and the result matrix $\hat{A}$ as consistent as possible, we minimized the following loss function:

$$L_1 = \left\|\tilde{A} - \hat{A}\right\|^2 = \sum_i \sum_j \left(\tilde{\mathrm{A}}(i,j) - \hat{A}(i,j)\right)^2. \tag{4}$$

Through this graph convolutional autoencoder, we deeply mined the potential relationships between the nodes in dtNet.

2.3.3 Adversarial Model

It is assumed that the feature vector z_i of a drug or target obeys a Gaussian distribution $z_i \sim q(z_i)$. To improve the robustness of the feature vectors obtained by the encoder, we introduced a generative adversarial network (GAN) to make the low-dimensional feature vectors of the drugs or targets better fit a Gaussian distribution. A multi-layer perceptron (MLP)-based discriminator D was constructed to determine whether the input vector of D is from generator G or a random vector, which was sampled from a Gaussian distribution $z_i^{'} \sim p(z_i^{'})$.

We first performed random sampling on a Gaussian distribution $z_i^{'} \sim p(z_i^{'})$ for $m{+}n$ times, and obtained a true sample set $Z' = \{z_1^{'}, z_2^{'}, z_3^{'} \ldots z_{m+n}^{'}\}$ obeying a Gaussian

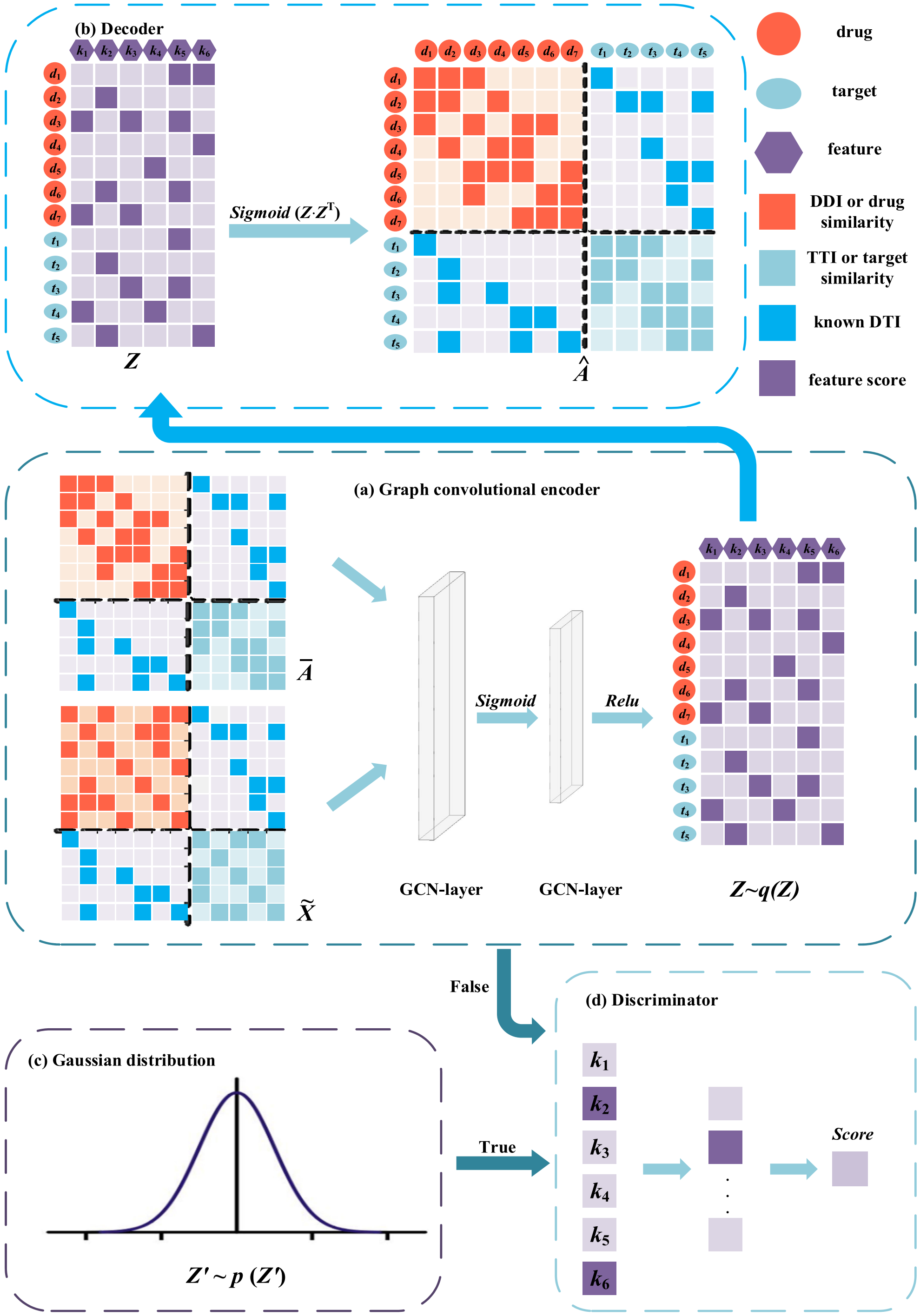


Fig. 2. The learning process of low-dimension feature vectors. (a) The graph convolutional encoder; (b) decoder and the reconstructed topological information matrix $\hat{A}$; (c) Gaussian distribution; and (d) the MLP-based discriminator.

distribution, where $z_i \epsilon R^k$. To avoid class imbalance in the input data of the discriminator, we randomly sampled a vectors in the embedding representation matrix Z and the true sample set Z' to construct the input matrix $X^D \epsilon R^{2a \times k}$ of the discriminator D. For each input vector x_i^D, the discriminator gives a score between 0 and 1 to determine whether the vector is from $Z' \sim p(Z')$ or generator G. The closer the score is to 1, the more likely x_i^D is to be sampled in a true Gaussian distribution, and vice versa. The score s_i for the input vector x_i^D obtained by discriminator D can be calculated as follows:

$$s_i = \phi_2\left(w_i^2 \cdot \phi_1\left(w_i^1 \cdot x_i^D + b_i^1\right) + b_i^2\right), \tag{5}$$

where w_i^1 and b_i^1 are the weight and bias vectors of the first fully connected layer, respectively, and w_i^2 and b_i^2 are for the second. $\phi_1(\cdot)$ is the relu activation function and $\phi_2(\cdot)$ is the sigmoid activation function. The loss L_2 of GAN was calculated as follows:

$$L_2 = -\frac{1}{2}\sum_i \left(E_{x_i \sim \mathrm{p(x)}}[\log s_i] + E_{x_i \sim \mathrm{q(x)}}[\log\left(1 - s_i\right)]\right). \tag{6}$$

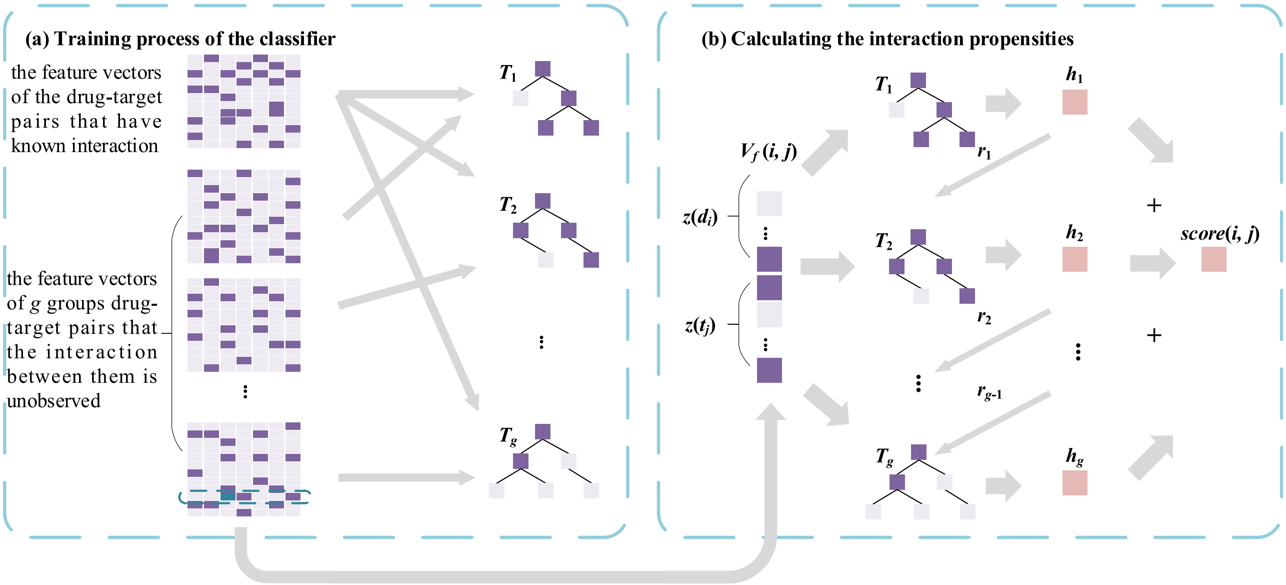

(a) Training process of the classifier
the feature vectors of the drug-target pairs that have known interaction
the feature vectors of g groups drug-target pairs that the interaction between them is unobserved
T1
T2
Tg
(b) Calculating the interaction propensities
V_f (i, j)
z(d_i)
z(t_j)
T1
T2
Tg
r1
r2
r_g-1
h1
h2
hg
+
+
score(i, j)

Fig. 3. Calculating the interaction propensities by LightGBM-based classifier.

The generator G wants to generate a feature vector that fits the Gaussian distribution to fool discriminator D, thus ensuring D scores it as high as possible. However, the purpose of discriminator D is to score the input vector from generator G as low as possible and score the input vector sampled in p as high as possible. Thus, the optimization goal of GAN can be defined as follows:

$$O_2 = \min_G \max_D \sum_i \left(E_{x_i \sim \mathrm{p(x)}}[\log s_i] + E_{x_i \sim \mathrm{q(x)}}[\log(1 - s_i)] \right). \quad (7)$$

2.3.4 Classifier Based on LightGBM

In our dataset, there is a serious class imbalance between the known (positive samples) and unknown DTIs (negative samples). The ratio between positive and negative samples is 1:120. For training traditional machine learning models such as KNN [32] and SVM [33], the same number of negative and positive samples were used. Many negative samples containing valuable information were abandoned, limiting the accuracy of the prediction results. To release the negative impact of class imbalance, we proposed an ensemble learning model based on LightGBM as a classifier. LightGBM can effectively release the effect of class imbalance by establishing multiple decision trees. It uses a different dataset of negative samples to train different trees, ensuring full utilization of the negative samples.

We received the drug and target feature vectors from the adversarial graph convolutional encoder. If we use $Z(d_i)$ to represent the feature vector d_i and use $Z(t_j)$ to represent the feature vector of t_j, the feature vector of drug-target pair (d_i, t_j) represented by $V_f(i,j)$ can be obtained by concatenating $Z(d_i)$ and $Z(t_j)$. Assuming that the ratio of positive and negative samples in the dataset is 1:g, g decision trees will be built and denoted by $T \ = \{T_1, T_2, T_3, \ldots, T_g\}$. For the first k decision trees, we can obtain k scores of $V_f(i,j)$ from these decision trees. Based on these scores and the label of $V_f(i,j)$, we can calculate a residual r_k, which will be used as the label of the k+1-th decision tree. r_k can be calculated as follows:

$$r_k = Y_{ij} \ - \sum_{t\,=\,1}^{k-1} T_t\big(V_f(i,j)\big), \quad (8)$$

where $T_t(V_f(i,j))$ is the score of the t-th decision tree. All negative samples were equally and randomly divided into g groups. Each decision tree T_k $(1 \leq k \leq g)$ was trained with a group of negative samples and all positive samples (as shown in Fig. 3(a)). Thus, in the training set of each decision tree, there were the same number of positive and negative samples. After training the model, we used all decision trees to score $V_f(i,j)$ and summarized the scores as the propensity of d_i for interacting with t_j. The interaction score of (d_i, t_j) can be defined as follows:

$$score\ (i,j) = \sum_{k=1}^{g} T_k\big(V_f(i,j)\big). \quad (9)$$

The higher the $score(i,j)$, the higher the possibility that d_i interacted with t_j. The matrix of the interaction score $\hat{Y} \in R^{m\times n}$ can be defined as follows:

$$\hat{Y}_{ij} = \ score(i,j). \quad (10)$$

The loss of GANDTI was evaluated by root-mean-square error. To improve the accuracy of the prediction results, the prediction scores of the positive samples were anticipated to be as high as possible, while the negative sample scores were as close to 0 as possible. Therefore, the prediction model can be optimized by Equation (11):

$$O_3 = \min\left(\sum_{i,j} \big(Y_{ij} - \hat{Y}_{ij}\big)^2 \right), \quad (11)$$

where Y_{ij} is the real interaction between d_i and t_j.

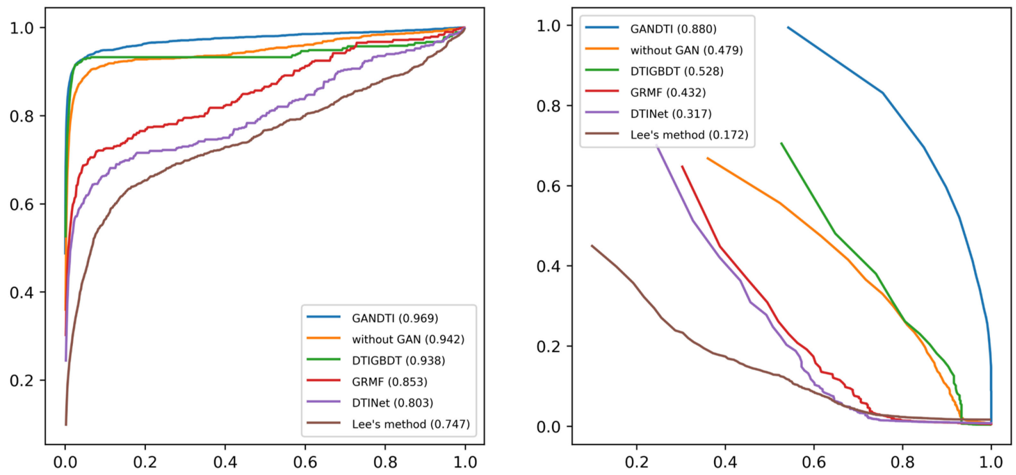


Fig. 4. ROC and P-R curves of different DTI prediction methods.

3 Results and Discussion

3.1 Performance Evaluation Metrics

A five-fold cross-validation [34] was used to evaluate the performance of the algorithm. The basic idea of cross-validation was to divide the original data sets into different groups. These groups were taken as training and testing sets in turn.

All known DTIs (positive samples) were randomly divided into five groups of equal size. The same operation was applied to unknown DTIs (negative samples). In each fold of validation, four groups of known and unknown DTIs were used to train the model, and the remaining DTIs were used for testing. Therefore, a total of 1,536 known DTIs and 184,320 unknown DTIs were used to train the model. The remaining 387 known DTIs and 46,533 unknown DTIs were regarded as the test set. After calculating the interaction scores of all drug targets using the prediction model, the samples were sorted by their prediction scores in descending order. The higher the ranking of the positive samples, the better the performance of the method.

For a given threshold $\boldsymbol{\delta}$, if the predicted score of a positive sample was greater than $\boldsymbol{\delta}$, it was taken as a true positive sample (TP). If the score of a positive sample was lower than $\boldsymbol{\delta}$, it was defined as a false negative sample (FN). If the score of a negative sample was greater than $\boldsymbol{\delta}$, it was considered a false positive sample (FP). If not, it was regarded as a true negative sample (TN). The receiver operating characteristic (ROC) curve [35] can be constructed by calculating the true positive rates (TPRs) and false positive rates (FPRs) under various $\boldsymbol{\delta}$. The TPRs and FPRs can be defined as follows:

$$TPR = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{TN + FP}. \tag{12}$$

The area under the ROC curve (AUC) was used to evaluate the performance of the prediction method [36]. By general consensus, the closer the AUC is to 1, the better the performance of the method. However, previous studies have shown that for data with class imbalance, the area under the P-R curve (AUPR) is a more informative metric [37]. We therefore also used the AUPR to evaluate our method, which was calculated by precision and recall. The precision and recall rates can be defined as follows:

$$Precision = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN}. \tag{13}$$

In addition, biologists usually select the top section of the prediction results for further validation through wet-lab experiments. The accuracy of the top k candidate targets predicted for each drug was therefore more important [38]. Hence, we also showed the recall rates of the top k ($k = 30$, $60 \ldots 240$) candidate drug-target pairs to reveal how many positive samples were successfully identified in the top k candidates.

3.2 Parameter Setting

For the weight matrix $W_1 \epsilon R^{(m+n)\times l}$ and $W_2 \epsilon R^{l\times k}$ of the encoder, the settings in GANDTI were $l = 500$ and $k = 200$. The number of samples a in the GAN was set to 900. We used PyTorch to train and optimize the neural network on a GPU (Nvidia GeForce RTX 2070) device. The epoch and learning rate of the neural network were set to 2000 and 0.005, respectively.

3.3 Comparison With Other Methods

To evaluate the performance of GANDTI, we compared it to several state-of-the-art DTI prediction methods, including DTIGBDT [21], GRMF [17], DTINet [18], and Lee's method [20]. For fairness of comparison, the hyperparameters in each model were set to the recommendations of the corresponding literature ($a = 0.4$, k $= 30$, $\lambda = 0.1$ for DTIGBDT; $\eta = 0.5$, $d = 0.1$, $t = 0.1$, $l = 2$ for GRMF; $r = 0.8$,$\lambda = 1$ for DTINet; and $r = 0.8$ for Lee's method). In particular, we compared our model to the one without GAN to demonstrate the efficiency of the GAN.

The ROC and P-R curves of each method are listed in Fig 4. GANDTI achieved the best performance (AUC $=$ 0.969, AUPR $=$ 0.880), obtaining 3.1 percent higher AUC and 35.2 percent higher AUPR values than the second model, DTIGBDT. DTIGBDT only extracted the path-based features for each drug-target pair, without learning their deeper features. Compared with GRMF and DTINet, the AUC for GANDTI was 11.6 and 16.6 percent higher than the other

TABLE 1
Results of Wilcoxon Test Between GANDTI and Other Methods Based on AUCs and AUPRs

	DTIGBDT	GRMF	DTINet	Lee's method	without GAN
p-values based on AUC	1.2052e-04	2.7691e-06	1.9377e-07	2.8530e-12	1.6186e-04
p-values based on AUPR	7.3248e-14	1.6052e-23	6.8440e-68	1.0169e-118	4.3746e-18

two methods, respectively. The AUPR for GANDTI was 44.8 and 56.3 percent higher than the GRMF and DTINet methods, respectively. This might be due to the existence of complex information between the nodes in the heterogeneous drug-target network. The shallow prediction models GRMF and DTINet based on matrix decomposition cannot capture deeper potential associations between the nodes. Lee's method achieved the worst performance, with AUC and AUPR values 22.2 and 70.8 percent lower than that of GANDTI. As KNN used the same quantities of negative and positive samples to train the model, most of the negative samples were not exploited. As for the model without GAN, GANDTI achieved 2.7 percent higher AUC and 40.1 percent higher AUPR. It indicates that the distribution of the feature vectors affects the accuracy of the prediction results. The superior performance of GANDTI can mainly be attributed to the method not only integrating the topological information and node attributes of the heterogeneous network, but also fully exploiting the negative samples of the dataset.

To verify whether the AUC and AUPR values of GANDTI were significantly superior to those of other methods, we performed a Wilcoxon test [39]. The statistical results listed in Table 1 show that GANDTI achieved significantly greater performance than all other methods at a p-value threshold of 0.05.

For the k-ranked targets, a higher recall rate corresponded with more positive samples being successfully identified [40]. The average recall rates across all tested drugs among the top k (k = 30,60,90...180) candidate targets were calculated, and the results are shown in Fig 5. The recall rates for GANDTI were superior to all other methods at various k values. GANDTI achieved 83.9, 89.4, and 93.7 percent positive samples in the top 30, 60, and 180, respectively. DTIGBDT achieved the second-best performance, and it yielded 74.5, 81.6, and 89.3 percent potential DTIs in the top 30, 60, and 180. GRMF demonstrated 70.7 percent in the top 30, 73.8 percent in the top 60, and 84.9 percent in the top 180. The performance of GRMF was slightly inferior to that of DTIGBDT. DTINet yielded 64.5 percent in the top 30, 69.6 percent in the top 60, and 77.2 percent in the top 180, and its recall rates were greater than those of Lee's method. Lee's method achieved the worst performance, and had only 50.8 percent in the top 30, 61.7 percent in the top 60, and 73.3 percent in the top 180.

3.4 Case Studies on Five Drugs

To demonstrate the ability of GANDTI to discover potential DTIs, we appiled case studies for five most connective drugs, namely *Quetiapine*, *Clozapine*, *Olanzapine*, *Aripiprazole*, and *Ziprasidone*. The top 10 ranked candidate targets for each drug are listed in Table 2. We consulted several reference databases and literature sources to support the prediction results of GANDTI.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database used to examine the drug-target interaction network [41]. Its data are mainly derived from published literature. Another database named DrugBank provides detailed drug data including drug-target interactions [29]. Universal Protein (UniProt) is a protein database that contains a large amount of information regarding the biological functions of proteins sourced from literature, such as the drugs' regulating effects on proteins [42]. As presented in Table 2, 19 candidate DTIs can be inferred by the KEGG database, 7 candidate interactions by the DrugBank, and 15 candidates by the UniProt, which means that these candidate targets may lead to a disease which is the indication of the drug. These items suggest that these drugs are likely to affect the expression of their candidate targets.

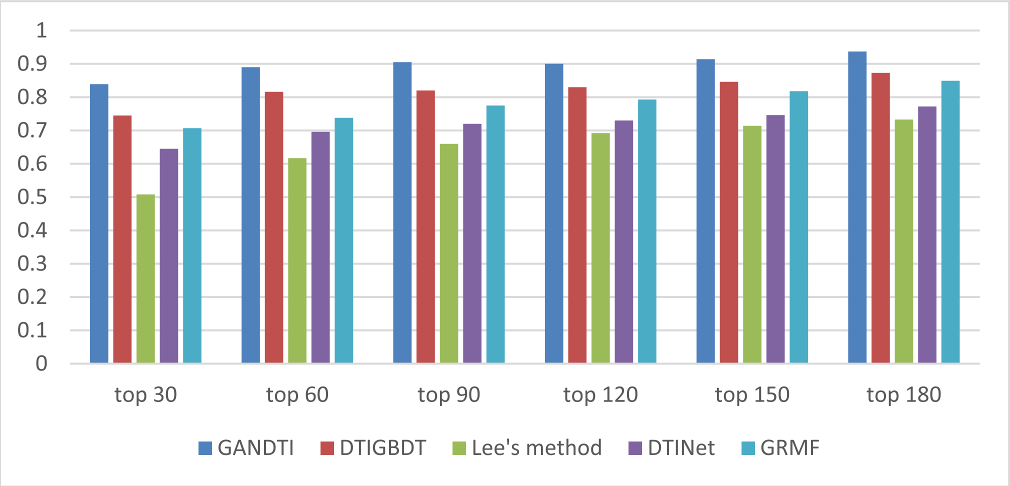


Fig. 5. Average recall rates across all tested drugs at different top k values.

TABLE 2
The 10 Top-Ranked Candidate Targets of Five Drugs

Quetiapine					
Rank	Target	Evidence	Rank	Target	Evidence
1	*HTR2B*	DrugBank	6	*SLC6A4*	Literature [44]
2	*ADRB2*	UniProt	7	*ADRB3*	UniProt
3	*HRH2*	KEGG	8	*HTR4*	KEGG
4	*ADRB1*	KEGG	9	*SLC6A2*	UniProt
5	*KCNMA1*	inferred by 1 literature	10	*GABRA1*	KEGG
Clozapine					
Rank	Target	Evidence	Rank	Target	Evidence
1	*DRD5*	KEGG	6	*GABRA1*	KEGG
2	*HRH2*	UniProt	7	*HTR1F*	UniProt
3	*ADRA1D*	KEGG	8	*ADRB2*	DrugBank
4	*HTR2B*	KEGG	9	*HTR4*	UniProt
5	*TSPO*	PhID	10	*GABRD*	KEGG
Olanzapine					
Rank	Target	Evidence	Rank	Target	Evidence
1	*ADRA1D*	KEGG	6	*GABRA3*	UniProt
2	*HRH2*	Uniport	7	*ADRB2*	KEGG
3	*HTR2B*	Uniport	8	*OPRM1*	DrugBank
4	*GABRA1*	Uniport	9	*GABRA4*	UniProt
5	*HTR4*	KEGG	10	*ABL1*	KEGG
Aripiprazole					
Rank	Target	Evidence	Rank	Target	Evidence
1	*ADRA1D*	Literature [45]	6	*HRH2*	UniProt
2	*SCN5A*	KEGG	7	*ADRB3*	DrugBank
3	*HTR2B*	PhID	8	*PTGS2*	KEGG
4	*ADRB2*	KEGG	9	*PDE4B*	UniProt
5	*HTR4*	KEGG	10	*OPRK1*	DrugBank
Ziprasidone					
Rank	Target	Evidence	Rank	Target	Evidence
1	*HRH2*	DrugBank	6	*SCN5A*	UniProt
2	*HTR2B*	PhID	7	*GABRR1*	UniProt
3	*ADRA1D*	Literature [46]	8	*GABRD*	KEGG
4	*GABRA1*	PhID	9	*GABRR2*	DrugBank
5	*GABRA3*	PhID	10	*TSPO*	KEGG

A database called PhID has been developed for network pharmacology research [43]. It contains real drug-target interaction information that was verified by the wet-lab experiments. Some candidate DTIs labeled with 'literature' were supported by some published literature. Five candidate DTIs in the table are supported by PhID and three candidates were reported by previous literature, indicating that there are indeed interactions between these drugs and their candidate targets, and that these have been confirmed experimentally.

In addition to the manually verified DTIs, the DrugBank database also contains some potential interactions inferred by literature. One candidate target of Quetiapine, *KCNMA1*, was contained in the inferred part of DrugBank, which suggests that the expression of *KCNMA1* is likely adjusted by *Quetiapine*. All candidate DTIs listed in Table 2 are supported by relevant databases or existing literature. This demonstrates the powerful ability of GANDTI to determine potential DTIs. Supplementary Table ST1 lists 30 high-quality candidate targets for each drug and their interaction scores.

4 Conclusions

A graph convolutional autoencoder and generative adversarial network-based method, GANDTI, was developed to predict novel DTIs. The graph convolutional autoencoder of GANDTI captured multiple types of intra-connections about drugs and targets, such as drug and target interactions and similarities. Meanwhile, it also captured the inter-connections between drugs and targets (known DTIs). Moreover, the low-dimension feature distribution of the drug and target nodes was regularized by a generative adversarial network, used to enhance the evidence of DTIs. The ensemble learning-based classifier LightGBM completely exploited all the negative samples, effectively counteracting the effect of class imbalance. The experimental results indicate that the performance of GANDTI is superior to all other methods tested here in terms of both AUCs and AUPRs. GANDTI is a more useful method for biologists as its top-ranking list contains more actual DTIs. Case studies on five drugs demonstrated the ability of GANDTI to discover potential DTIs. GANDTI is a powerful prioritization tool that provides biologists with reliable candidate DTIs, used for subsequent identification of actual DTIs with wet-lab experiments.

In the recent years, with the deepening of research, some non-coding RNAs, such as microRNAs and long non-coding RNAs, have been discovered that may affect gene expression and disease progression. Some studies have also shown that non-coding RNAs can be regarded as a new type of drug targets [47], [48], [49], [50], [51]. Therefore, our subsequent research may involve the introduction of information related to non-coding RNAs to assist the prediction of DTIs.

Acknowledgments

The work was supported by the Natural Science Foundation of China (61972135), the Natural Science Foundation of Heilongjiang Province (LH2019F049, LH2019A029), the China Postdoctoral Science Foundation (2019M650069), the Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHL-Q18104), the Fundamental Research Foundation of Universities in Heilongjiang Province for Technology Innovation (KJCX201805), the Foundation of Graduate Innovative Research (YJSCX2019-070HLJU), and the Fundamental Research Foundation of Universities in Heilongjiang Province for Youth Innovation Team (RCYJTD201805).

References

[1] R. S. Olayan, H. Ashoor, and V. B. Bajic, "DDR: Efficient computational method to predict drug-target interactions using graph mining and machine learning approaches," *Bioinformatics*, vol. 373, no. 2054, 2017, Art. no. 20140419.

[2] J. P. Overington, A. L. Bissan, and A. L. Hopkins, "How many drug targets are there?" *Nature Rev. Drug Discov.*, vol. 5, no. 12, pp. 993–996, 2006.

[3] R. Santos *et al.*, "A comprehensive map of molecular drug targets," *Nature Rev. Drug Discov.*, vol. 16, no. 1, pp. 19–34, 2017.

[4] A. Cichonska, J. Rousu, and T. Aittokallio, "Identification of drug candidates and repurposing opportunities through compound–target interaction networks," *Expert Opinion Drug Discov.*, vol. 10, no. 12, pp. 1333–1345, 2015.

[5] X. Chen *et al.*, "Drug-target interaction prediction: Databases, web servers and computational models," *Brief Bioinf.*, vol. 17, no. 4, pp. 696–712, 2016.

[6] S. Zheng, Y. Li, S. Chen, J. Xu, and Y. Yang, "Predicting drug protein interaction using quasi-visual question answering system," *Nature Mach. Intell.*, vol. 2, pp. 134–140, 2019.

[7] Y. A. Huang, Z.-H. You, and X. Chen, "A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences," vol. 19, no. 5, pp. 468–478, 2018.

[8] L. Wang *et al.*, "RFDT: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information," *Current Protein Peptide Sci.*, vol. 19, no. 5, pp. 445–454, 2018.

[9] Z. Qi, H. Yu, M. Ji, Y. Zhao, and X. Chen, "Computational model development of drug-target interaction prediction: A review," *Current Protein Peptide Sci.*, vol. 20, no. 6, pp. 492–494, 2019.

[10] A. C. Cheng *et al.*, "Structure-based maximal affinity model predicts small-molecule druggability," *Nature BioTechnol.*, vol. 25, no. 1, pp. 71–75, 2007.

[11] G. M. Morris *et al.*, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2010.

[12] B. R. Donald, *Algorithms in Structural Molecular Biology*, Cambridge, MA, USA: MIT Press, 2011.

[13] M. J. Keiser *et al.*, "Relating protein pharmacology by ligand chemistry," *Nature BioTechnol.*, vol. 25, no. 2, pp. 197–206, 2007.

[14] J. Li *et al.*, "A survey of current trends in computational drug repositioning," *Briefings Bioinf.*, vol. 17, no. 1, pp. 2–12, 2016.

[15] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artif. Intell. Medicine*, vol. 77, no. C, pp. 53–63, 2017.

[16] X. Chen, M. X. Liu, and G. Y. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Mol. Biosyst.*, vol. 8, no. 7, pp. 1970–1978, 2012.

[17] A. Ezzat, P. Zhao, M. Wu, X. Li, and C. K. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 646–656, May/Jun. 2017.

[18] Y. Luo *et al.*, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 573.

[19] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[20] I. Lee and H. Nam, "Identification of drug-target interaction by a random walk with restart method on an interactome network," *BMC Bioinformatics*, vol. 19, no. 8, 2018, Art. no. 208.

[21] P. Xuan *et al.*, "Gradient boosting decision tree-based method for predicting interactions between target genes and drugs," *Front. Genetics*, vol. 10, no. 459, 2019, Art. no. 459.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017, pp. 1–14.

[23] K. F. Wang, C. Gou, Y. J. Duan, Y. L. Lin, and F. Y. Wang, "Generative adversarial networks: The state of the art and beyond," *Acta Automatica Sinica*, vol. 43, no. 3, pp. 321–332, 2017.

[24] S. Pan, R. Hu, S. F. Fung, G. Long, and C. Zhang, "Learning graph embedding with adversarial training methods," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2168–2275, Sep. 2019.

[25] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[26] N. C. Oza, "Online ensemble learning," in *Proc. 17th Nat. Conf. Artif. Intell. 12th Conf. Innovative Appl. Artif. Intell.*, 2000, Art. no. 1.

[27] M. A. Friedl, and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environment*, vol. 61, no. 3, pp. 399–409, 1997.

[28] T. Keshava Prasad *et al.*, "Human protein reference database—2009 update," *Nucleic Acids Res.*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

[29] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2017.

[30] I. Francesco *et al.*, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 8, pp. 14621–14626, 2010.

[31] W. Wenhui, Y. Sen, Z. Xiang, and L. Jing, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[32] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Inform.*, vol. 12, no. 1, pp. 90–108, 2016.

[33] A. Abdiansah and R. Wardoyo, "Time complexity analysis of support vector machines (SVM) in LibSVM," *Int. J. Comput. Appl.*, vol. 128, no. 3, pp. 975–8887, 2015.

[34] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.

[35] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, 1993.

[36] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[37] P. Xuan, T. Shen, X. Wang, T. Zhang, and W. Zhang, "Inferring disease-associated microRNAs in heterogeneous networks with node attributes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2018. to be published, doi 10.1109/TCBB.2018.2872574

[38] P. Xuan, Y. Dong, Y. Guo, T. Zhang, and Y. Liu, "Dual convolutional neural network based method for predicting disease-related miRNAs," *Int. J. Mol. Sci.*, vol. 19, no. 12, pp. 3732, 2018.

[39] P. Xuan, Y. Ye, T. Zhang, L. Zhao, and C. Sun, "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations," *Cells*, vol. 8, no. 7, 2019, Art. no. 705.

[40] P. Xuan *et al.*, "Drug repositioning through integration of prior knowledge and projections of drugs and diseases," *Bioinformatics*, vol. 35, no. 20, pp. 4108–4119, 2019.

[41] M. Kanehisa, and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[42] A. Rolf *et al.*, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 32, no. Database, pp. D115–D119, 2004.

[43] Z. Deng, W. Tu, Z. Deng, and Q.-N. Hu, "PhID: An open-access integrated pharmacology interactions database for drugs, targets, diseases, genes, side-effects, and pathways," *J. Chem. Inf. Model.*, vol. 57, no. 10, pp. 2395–2400, 2017.

[44] H. Sugawara *et al.*, "Effects of quetiapine on DNA methylation in neuroblastoma cells," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 56, pp. 117–121, 2015.

[45] A. Botzer, Y. Finkelstein, E. Grossman, J. Moult, and R. Unger, "Iatrogenic hypertension: A bioinformatic analysis," *Pharmacogenomics J.*, vol. 19, no. 4, pp. 337–346, 2019.

[46] J. Sun and Z. Zhao, "Pathway-assisted investigation of atypical antipsychotic drugs and serotonin receptors in schizophrenia," in *Proc. Biomed. Sci. Eng. Conf.*, 2010, pp. 1–4.

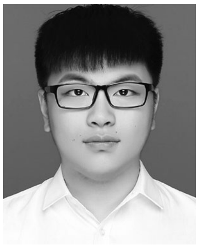
[47] C. C. Wang, X. Chen, J. Qu, Y. Z. Sun, and J. Q. Li, "RFSMMA: A new computational model to identify and prioritize potential small molecule-MiRNA associations," *J. Chem. Inf. Model*, vol. 59, no. 4, pp. 1668–1679, 2019.

[48] X. Chen, N. N. Guan, Y. Z. Sun, J. Q. Li, and J. Qu, "MicroRNA-small molecule association identification: From experimental results to computational models," *Brief Bioinf.*, vol. 21, no. 1, pp. 47–61, 2018.

[49] J. Yin, X. Chen, C. C. Wang, Y. Zhao, and Y. Z. Sun, "Prediction of small molecule-microRNA associations by sparse learning and heterogeneous graph inference," *Mol. Pharm.*, vol. 16, no. 7, pp. 3157–3166, 2019.

[50] J. Qu, X. Chen, Y. Z. Sun, J. Q. Li, and Z. Ming, "Inferring potential small molecule-miRNA association based on triple layer heterogeneous network," *J. Cheminformatics*, vol. 10, no. 1, 2018, Art. no. 30.

[51] J. Qu *et al.*, "In silico prediction of small molecule-miRNA associations based on the HeteSim algorithm," *Mol. Ther. - Nucleic Acids*, vol. 14, pp. 274–286, 2019.



Chang Sun received the bachelor's degree in computer science and technology from Heilongjiang University. He is currently working toward the master's degree in the School of Computer Science and Technology at Heilongjiang University, Harbin, China. His research interests include complex network analysis and deep learning.

Ping Xuan received the PhD degree from the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2012. She is currently a professor of computer science at Heilongjiang University, Harbin, China. She is also a postdoctoral research fellow with the College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China. She visited Washington University in St. Louis, St. Louis, USA, as a visiting scholar from 2015 to 2016. Her current research interests include computational biology, deep learning, and natural language processing.

Tiangang Zhang received the PhD from the Graduate School of Engineering, the University of Tokyo, Tokyo, Japan, in 2016. He is an associate professor of the Department of Mathematical Science, Heilongjiang University, Harbin, China. His current research interests include complex network analysis and computational fluid dynamics.

Yilin Ye received the bachelor's degree in computer science and technology from Heilongjiang University, he is currently working toward the master's degree in the School of Computer Science and Technology at Heilongjiang University, Harbin, China. His research interests include complex network analysis and deep learning.