RESEARCH PAPER

Relational Topology-based Heterogeneous Network Embedding for Predicting Drug-Target Interactions

Linlin Zhang[1], Chunping Ouyang[1,2†], Fuyu Hu[1†], Yongbin Liu[1,2], Zheng Gao[3]

[1]School of Computer, University of South China, Hengyang,Hunan, 421001, China

[2]Hunan Medical Big Data International Sci.&Tech, Innovation Cooperation Base

[3]Department of Information and Library Science, Indiana University Bloomington Woodlawn Avenue, IN 47408, Bloomington, America

Keywords: Link prediction; Heterogeneous information network; Drug-target interaction; Network embedding; Feature representation

Citation: Zhang, L.L., Ouyang, C.P., Hu, F.Y., et al.: Relational Topology-based Heterogeneous Network Embedding for Predicting Drug-Target Interactions. Data Intelligence 5(2), 475-493 (2023). doi: 10.1162/dint_a_00149
Received: Dec. 20, 2021; Revised: Apr. 15, 2022; Accepted: May 19, 2022

ABSTRACT

Predicting interactions between drugs and target proteins has become an essential task in the drug discovery process. Although the method of validation via wet-lab experiments has become available, experimental methods for drug-target interaction (DTI) identification remain either time consuming or heavily dependent on domain expertise. Therefore, various computational models have been proposed to predict possible interactions between drugs and target proteins. However, most prediction methods do not consider the topological structures characteristics of the relationship. In this paper, we propose a relational topology-based heterogeneous network embedding method to predict drug-target interactions, abbreviated as RTHNE_DTI. We first construct a heterogeneous information network based on the interaction between different types of nodes, to enhance the ability of association discovery by fully considering the topology of the network. Then drug and target protein nodes can be represented by the other types of nodes. According to the different topological structure of the relationship between the nodes, we divide the relationship in the heterogeneous network into two categories and model them separately. Extensive experiments on the real-world drug datasets, RTHNE_DTI produces high efficiency and outperforms other state-of-the-art methods. RTHNE_DTI can be further used to predict the interaction between unknown interaction drug-target pairs.

† Corresponding author: Chunping Ouyang (Email: ouyangcp@126.com; ORCID: 0000-0002-2154-0079); Fuyu Hu (Email: fyhu1124@163.com; ORCID: 0000-0002-1004-6913).

1. INTRODUCTION

The prediction of drug-target interactions (DTIs) is the process of uncovering potential drug-target interactions that are currently undiscovered through information on the structural characteristics of drugs and targets, as well as known drug-target relationships and drug-drug relationships. Prediction of DTIs is the key to developing new drugs. It plays an important role in the study of drug toxicity and side effects and the treatment of diseases. However, traditional methods based on large-scale biological experiments usually take several years and are often very expensive [1]. Therefore, the prediction of DTIs has received great attention and computer-aided drug discovery has become a trend. In recent years, with the rapid development of computer technology and the accumulation of large amounts of medical data, computational solutions have shown feasible and reasonable in discovering new drug-target interactions (DTIs) in the age of big data, methods such as machine learning and data mining have been widely used to solve various complex problems in the field of biomedicine [2, 3, 4]. Currently, there are three types of prediction approaches in computational-aided drug discovery, namely Attribute-based calculation methods [5], deep learning-based [6, 7], and network-based [8, 9] methods.

In the past years, several approaches have focused on prediction methods based on similarity calculation using the attributes of the drug and the target protein. The prediction method of similarity calculation usually depends on the attributes of the drug and the target protein, such as [5, 10]. Nevertheless, most of these methods use the chemical structure and protein sequence of the drug. In public data sets, it is often difficult to obtain the protein sequence and chemical information of many polymers. Deep learning has advanced computational modeling of DTI by offering an increased feature extraction power in drug and protein data. Such as [6, 7] are the prediction models of DTIs by deep learning. However, these deep learning methods do not consider the network topology information of drugs and target proteins, and the prediction results are often unsatisfactory. In recent years, network-based approaches have demonstrated great advantages compared to similarity-based methods. Especially heterogeneous network-based methods achieved good results in DTI prediction by considering a wide variety of topological information and the complex interaction relationship of heterogeneous data. Zeng et al. [9] constructed a heterogeneous network that covers the network profile and attributes information between drugs, target proteins, and diseases. An arbitrary-order approximate embedded deep forest method is used to predict DTI. The relationship between different nodes in a heterogeneous graph is different.

In this paper, we study the interaction relationship network between drugs and target proteins, and classify all relationships in the heterogeneous network into two categories and model them separately according to the characteristics of the topology of the relationships between nodes. This results in a topology-based RTHNE_DTI heterogeneous network embedding method that can be used to predict unknown drug-target protein interactions. Specifically, our main contributions are as follows:

- We proposed a heterogeneous network representation learning method named "RTHNE_DTI" to predict DTI, it learns the distributed representation of nodes by embedding heterogeneous network into low-dimensional spaces, which use the network topology information fully. On the other hand, we apply the method of heterogeneous network representation learning to drug-target interaction prediction, which achieves a more rapid and effective use of medical data, thereby significantly improving prediction accuracy.

- The traditional heterogeneous network representation learning method uses a uniform model to deal with all relationships. However, different node-node relationship represents different drug characteristics. In hence, we divide the relationship in a heterogeneous network into two types: Affiliation relationship and Peer relationship, and we design different models to represent them, which can better capture the rich feature information between nodes.
- In general, the prediction of drug-target interaction is carried out on the labeled network (Some drug target relationship pairs with known interactions were added to the training set). However, our model can also achieve good prediction results on the unlabeled network. This solves the problem of insufficient drug labeling data and low prediction accuracy.
- We conduct different experiments using real drug data set and compare with other predictive models, and the results show that RTHNE_DTI has the best predictive performance.

2. RELATED WORK

Computational methods in DTIs prediction have gained more attention because carrying out a biochemical experiment on a large scale is costly and time-consuming. The early computational methods mainly focused on similarity calculation between the attributes of the drug and the target protein. Then deep learning methods for computational prediction of DTI have become more popular in recent years. The rapid development of deep learning provides an effective way to predict DTI, especially for large-scale data prediction tasks. Mayr et al. [11] compared several deep learning methods with other machine learning and target prediction methods on large-scale drug discovery datasets and concluded that the deep learning method has the best prediction performance. Lee et al. [12] predicted DTIs through convolutional neural networks (CNNs) on original protein sequences. In a study called DeepDTA, Ozturk et al. [6] proposed a deep-learning- based model to predict the binding affinity between drugs and targets, CNNs were mainly used to model protein sequences and compound 1D representations. These methods are generally dependent on drug and target attributes.

In recent years, heterogeneous network representation learning has become a hot topic of current research and has good performance in link prediction [13]. Although heterogeneous network representation learning methods have been widely used for link prediction in social networks with good results. Most previous studies on networks have been based on homogeneous networks. Specifically, the nodes in the network are of the same type. With the development of network representation learning, in order to model the heterogeneity of networks, some have tried heterogeneous network representation learning. For example, Shang et al. [14] proposed a framework, ESim, which uses random wandering based on matching paths to generate sequences of nodes to optimize the similarity between multiple points. Fu et al. [15] proposed a heterogeneous information network representation, HIN2vec, which differs from many previous works based on skip gram language models in that the core of HIN2V ec is a neural network model that learns the representation of nodes and relations (meta-paths) in the network. Han et al. [16] proposed an aspect-level collaborative filtering model based on neural networks. In their model, they extract similarity matrices of different aspect levels of nodes through different meta-paths and feed these matrices into a deep neural network designed to learn aspect-level potential factors. These methods are commonly used in social networks, scholar networks, etc.

Therefore, prediction methods based on network topology are also widely used in DTI relationship prediction. There are various networks in practice, such as social networks [17], citation networks [18], and biological information networks [19]. And some interesting research works on network analysis have attracted increasing attention. Particularly, link prediction is one of the hot spot tasks of network analysis. Currently, most network-based DTI prediction is based on machine learning [8]. Wang et al. transformed new DTI prediction problems into a two-layer graphical model named the restricted Boltzmann machine (RBM). Wan et al. [20] developed a new nonlinear end-to-end learning model, called NeoDTI, which integrates different heterogeneous information of drugs and targets, and learned the representation of drugs and targets to predict DTIs. however, note that these approaches have the disadvantage of treating all node relationships in the heterogeneous network equally and they may not work when chemical pathways and protein interactions are unknown. Table 1 shows the categorization results.

Therefore, we apply a heterogeneous network embedding method to predict DTIs, called RTHNE_DTI. the biggest advantage of this method is that it can fully take into account the characteristics of different node relationships in the network and is modeled for that feature.

Table 1. An overview and comparison of related reviews.

	Attribute based calculation methods	Network topology-based methods	Deep learning-based methods
Y. Yamanishi [5]	√		√
X. Zheng [10]	√		
H. Öztürk [6]	√		√
Y.-B. Wang [7]	√		√
A. Mayr [11]	√		√
I. Lee [12]	√		√
X. Chen [8]		√	
X. Zeng [9]		√	
F. Wan [20]		√	√

3. PROBLEM DEFINITION

In this section, we introduce some basic definitions of heterogeneous network embedding to predict DTIs.

Definition 1: Heterogeneous Network (HN).

A Heterogeneous Network is defined as a Graph $G = (V, E, A, \emptyset, \psi)$, where V represents the set of nodes, $E \subseteq V \times V$ represents the set of edges. $\emptyset$ and ψ are the type mapping functions of nodes and edges, respectively, where $\emptyset = V \rightarrow N$ and ψ: $E \rightarrow R$. Here N and R are the type sets of nodes and edges, respectively. $A = N \cup R$, and while $|N| + |R| > 2$, the network is called a heterogeneous network; otherwise it is a homogeneous network.

Definition 2: Meta-path.

In a heterogeneous network, the meta-path P is a sequence of node types n_1, n_2, …, n_m and edge types r_1, r_2, …, r_{m-1}, in the form of:

$$P = n_1 \xrightarrow{r_1} n_2 \ldots \xrightarrow{r_{m-1}} n_m \tag{1}$$

Definition 3: Heterogeneous Network Embedding.

Given a heterogeneous network G, the heterogeneous network embedding learns a low-dimensional vector $E_v \in \mathbb{R}_d$ for each vertex $v \in V$ by a mapping function f: $V \rightarrow \mathbb{R}^d$, in which $d \ll |V|$ is the dimension of the representation space.

4. THE PROPOSED APPROACH

4.1 Overall Framework

As shown in Figure 1, (a) is a heterogeneous network constructed by five types of nodes (drug, target protein, disease, side-effect, action). In this network, there are not only simple relationships, such as D-D but also compound relationships, such as D-P-Di. In (b), we divide all relationships into two categories according to the relationship topology and model them separately. Finally, we apply the model to different scenarios to verify the performance of our model.

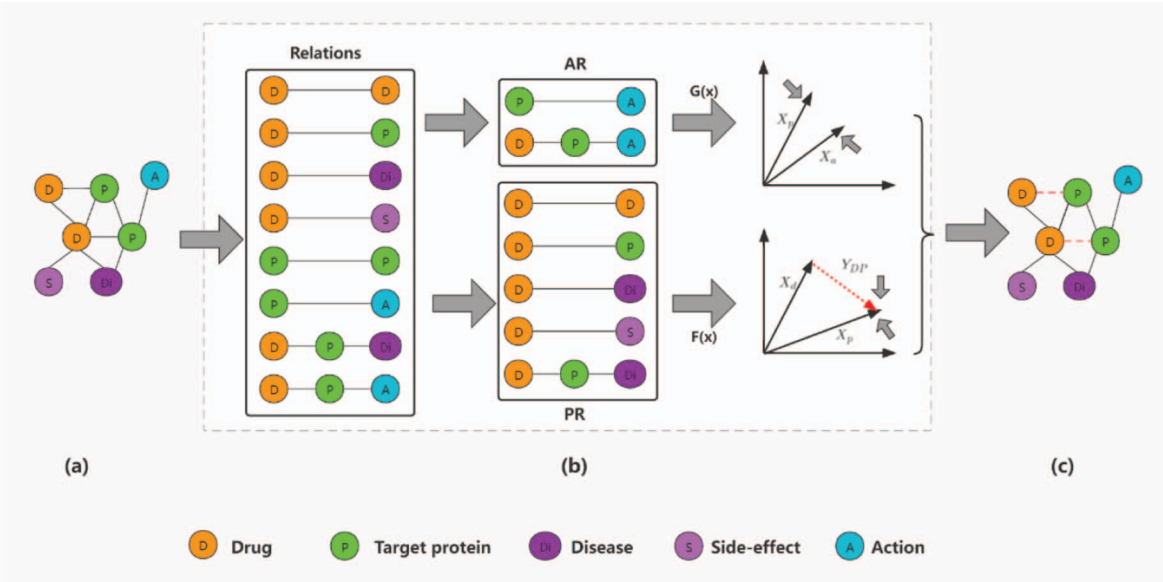


Figure 1. Structure of the RTHNE_DTI model. (a) A heterogeneous network of five classes of nodes (drug, target protein, disease, side effect and action) interacting with each other was constructed. (b) All relationships are classified into PR relationships and AR relationships. Then G(x) and F(x) are calculated respectively according to the characteristics of the two relationships during the training process, and the training weights are updated continuously update to obtain the node embedding results. (c) The embedding results are decoded to predict DTI.

4.2 Affiliation Relationship and Peer Relationship

In studying the data sets associated with the prediction of drug and target protein interactions, we found that not all relationship pairs had an equal number of nodes of the two types of connections, and some relationship pairs had a significantly different number of nodes of the two types of connections, as shown in figure 2.

Our study of DrugBank found that the types of action of proteins are very few, only 47, but the variety of proteins is wide. Hence, their relationship network looks like an action-centered network spreading outward. As shown in Figure 2 (a). However, most of the relationships in the drug data set are like drugs and proteins. The two types of nodes do not differ greatly in number, so they form a well-balanced network. As shown in Figure 2 (b).

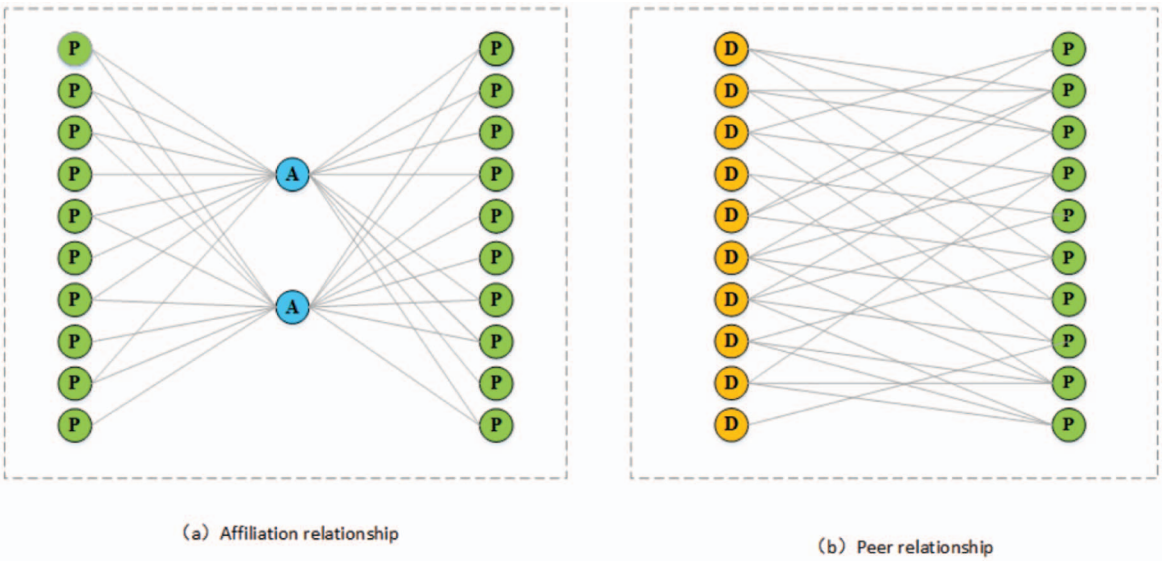


Figure 2. Different relationship topology. Node relationships in heterogeneous networks are classified into two categories based on the difference topology of nodes in the heterogeneous network. (a) Affiliation relationship with unbalanced distribution of nodes in network relationships; (b) peer relationship with balanced distribution of nodes in network relationships.

To fully use the topology characteristics of heterogeneous networks (HNs), we study the topological features of different relationships in a heterogeneous network. In the network, the degree of a node can well reflect the topological structure characteristics of the network [21]. In general, the degree of a node refers to the number of edges associated with the node. In order to explore the difference between the topological structures of different relationships in HN, we used the degree-based measure $D(e)$ for calculation:

$$D(e) = \frac{\max\left(\bar{d}n_u, \bar{d}n_v\right)}{\min\left(\bar{d}_{n_u}, \bar{d}_{n_v}\right)} \tag{2}$$

Where n_u, n_v represent the node type of nodes *u, v* in a relation tuple (*u, v, e*), $\overline{d}_{n_u}$ and $\overline{d}_{n_v}$ are the average degrees for n_u and n_v, respectively. It is worth mentioning that $D(e) \geq 1$. Here, a greater $D(e)$ value indicates that the topology of the two types of connected nodes is not identical, where one side is biased to the other? That is, nodes with a high $D(e)$ value show a stronger affiliation relationship (AR) between them, and nodes connected by this relationship have more similar attributes [22]. However, for smaller $D(e)$ values, it shows that the topology of the two types of connected nodes is peer, which we named the peer relationship (PR).

In order to ensure the accuracy of the results, we compare these relationships from the perspective of sparseness, annotated as *S(e)*, so as to discover the differences in the network structure of different relationships. We define *S(e)* as follows:

$$S(e) = \frac{N_r}{N_{n_u} \times N_{n_v}} \tag{3}$$

In the above formula, N_r is the number of edges in type *r*. In addition, N_{n_u} and N_{n_v} are the number of nodes of types n_u and n_v, respectively. It should be emphasized that in this way, these relationships can also be consistently divided into two categories, PR and AR.

Table 2. Statistical analysis of dataset.

Nodes	Num of Nodes	Relations	Num of Relations	Avg.D_u	Avg.D_v	$D(e)$	$S(e)$	Relationship type
Drug	708	D-D	10036	14.18	14.18	1.00	0.02002	PR
Target	1512	D-P	1923	2.72	1.27	2.14	0.00180	PR
Disease	5603	P-P	7360	4.87	4.87	1.00	0.00322	PR
Side-effect	4192	D-Di	199214	281.38	35.55	7.91	0.05022	PR
Action	47	D-S	80164	113.23	19.12	5.92	0.02701	PR
		P-D	1596745	1056.05	284.98	3.71	0.18848	PR
		P-A	2295	1.52	48.83	32.17	0.03229	AR

We conducted a comprehensive analysis of the obtained data according to the above indicators, as shown in Table 2.

4.3 Different models for PR and AR

To respect their different characteristics, we need to design different model treatments for them separately. Here, for two nodes connected by a PR relationship, there is a strong interactive relationship, and their topology structure is very similar. The nodes themselves contain rich structural information between two nodes, so we model the PR as a transition between nodes in a low-dimensional vector space.

In addition, for relation type AR, Euclidean distance is used as the calculation to measure the proximity of interacting nodes in low-dimensional space. It should be noted that the calculation methods we use for the two relationships are very consistent mathematically [23]. We use the Euclidean distance method for the AR mainly for the following reasons. First, the nodes connected by this relationship share the same

attributes [24], so the nodes connected by the AR can be directly approached in the vector space, which is consistent with the Euclidean distance optimization [25]. Second, the purpose of the heterogeneous network representation is to preserve the structural characteristics of the high-dimensional network. The Euclidean method satisfies the condition of triangular inequality [26], which ensures that the first-order and second-order similarities of the nodes remain unchanged.

Translation-based distance for peer relations. Through the study of Table 2, we found that in the heterogeneous drug network constructed, most of the relationships are peer-to-peer. Specifically, a drug acts on multiple diseases, and a disease can also be treated by multiple drugs. And the number of drug nodes and disease nodes differs very little. Peer relationships show powerful interactions between nodes with peer-to-peer structure. For the calculation of the score function of PR, we first give a PR-type relationship tuple (a, r, b), where $r \in R_{PR}$ has a weight of $w_{a,b}$. Then for the embedding of nodes a and b, we define them as P_a and P_b respectively. In addition, we annotate the embedding of relation r as Q_r. The final definition is as follows:

$$f(a,b) = w_{a,b}P_a + Q_r - P_b \tag{4}$$

For the relationship tuples $(a, r, b) \in T_{PR}$ whose relationship is PR in the heterogeneous network, the margin-based loss function [23] is defined as follows:

$$L_{PR} = \sum_{r \in PR(a,r,b) \in T_{PR}} \sum_{(a',r,b') \in T'_{PR}} max\left[0, \gamma + f(a,b) - f'_{(a,b)}\right] \tag{5}$$

In the above formula, T_{PR} represents the positive sample set in the PR triplet, and T'_{PR} is the negative sample set. $\gamma > 0$ represents a margin hyperparameter.

Euclidean distance for affiliation relations. Only the target protein and its action type belong to the AR relationship in the heterogeneous network we constructed. Specifically, the types of protein nodes and action nodes vary greatly in number. The nodes with this relationship can be directly approached in the vector space, so we use Euclidean distance to calculate the proximity between two nodes. Given a set of triples (m, i, n) with relationship type AR, where $i \in R_i$ represents the action relationship between nodes m and n. Its weight is defined as $w_{m,n}$ and the form is as follows:

$$g_{(m,n)} = w_{m,n}P_m - P_{n2}^2 \tag{6}$$

Similar to the above formula, P_m and P_n are the embedding of nodes m and n, respectively. $g(m, n)$ is to calculate the distance between m and n in a low-dimensional space. To ensure that the nodes connected by the AR relationship are closer, we minimize $g(m, n)$ as much as possible, therefore we define the margin-based loss function as:

$$L_{AR} = \sum_{r \in AR(m,i,n) \in T_{AR}} \sum_{(m',i,n') \in T'_{AR}} max\left[0, \gamma + g(m,n) - g'_{(m,n)}\right] \tag{7}$$

As before, T_{AR} and T'_{AR} are the positive and negative examples in the AR relationship, respectively.

4.4 Conjunctive Model

We have divided the node relationships of network into two categories based on the node distribution differences in the heterogeneous networks. Here, relations with unbalanced node distribution are Affiliation Relations (AR) and relations with balanced node distribution are Peer Relations (PR). We first initialize all the node embedding. Then for the node pairs of PR relations, we calculate the loss based on translation model, and for the node pairs of AR relations, we calculate the loss based on Euclidean Distance. Finally, we fuse the two losses and continuously modify the embedding according to the final loss value. Specifically, we make the model more complete by fusing the loss of the two models to jointly update the weight of node embedding. The pseudo code is shown in Algorithm 1.

Algorithm 1 : RTHNE_DTI model algorithm

Input: PR-type relationship tuple *(a,r,b)*,AR-type relationship tuple *(m,i,n)*
Output: W, P_m, P_n

Initialize W as training weights, *epochs* as training times, bs_{PR} as the batch_size of PR relations, bs_{AR} as the batchbs_size of AR relations
Initialization hyperparameter γ
for i=1,2,...epochs **do**
    for j=1,2,...bs$_{PR}$ **do**
        P_a, P_b=nn.embedding(a,r,b)
        $f(a,b) \leftarrow W||P_a+Q_r-P_b||$
        $loss \leftarrow \max[0, \gamma + f(a,b) - f'(a,b)]$
        update W with loss
    end for
    for k=1,2,...bs$_{AR}$ **do**
        P_m, P_n=nn.embedding(m,i,n)
        $g(m,n) \leftarrow W||P_m - P_n||_2^2$
        $loss \leftarrow \max[0, \gamma + g(m,n) - g'(m,n)]$
        update W with loss
    end for
end for
return W, P_m, P_n

5. EXPERIMENTS AND ANALYSIS

5.1 Datasets and Experimental Setup

In this paper, the data set we used to construct the heterogeneous network includes the node type set *V* = {drug, target, diseases, side-effects, action}, the relationship type set *R* = {drug-drug, drug-target, drug-diseases, drug-side-effects, target-target, target-diseases, target-action}. The data sources we used are as follows:

DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug-target interaction, target-action information, and drug-drug interaction information. We use the DrugBank version 3.0 and DrugBank version 5.1.6. [27].

HPDR (Human Protein Reference Database) contains manually curated scientific information pertaining to the biology of most human proteins and the data of protein interactions extracted from the HPRD database Release 9 [28].

CTD (Comparative Toxicogenomic Database) is a public website and research tool that provides four types of core data: chemical-gene interactions, chemical-disease associations, gene-disease associations, and chemical-phenotype associations. The drug-disease association and protein-disease association used in this paper were extracted from CTD [29].

SIDER database contains information about marketed drugs and their adverse reaction records. In this paper, the drug-side-effects interactions were extracted from SIDER database Version 2 [30].

We obtained data from the above four sources, and after data preprocessing, we finally got 708 drugs, 1, 512 target proteins, 5603 diseases, 47 actions, and 4, 192 side effects. Some descriptive statistics of the dataset are shown in Table 2.

RTHNE_DTI has three parameters: embedding dimension d, the margin γ, and α, we set $\gamma = 1$, and $\alpha = 0.01$. To study the influence of different dimensions on our model, we explored parameter d. As shown in figure 3, we can see that when the dimension is 300, the predicted AUC value is the highest. So we set d=100 in the experiment.

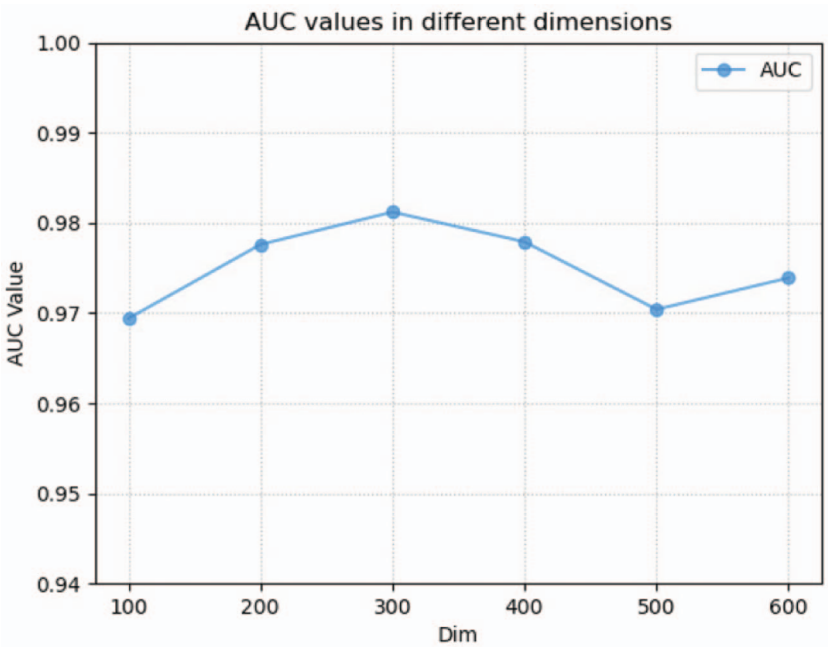


Figure 3. Parameter Analysis. Show the dimension value is when the best prediction result achieved.

In this paper, we conducted experiments under four tasks. In order to verify that dividing the relationship into AR and PR can effectively improve the prediction performance, we conducted two experiments on (1) the prediction performance of labeled networks based on PR relations only, and (2) the prediction performance of labeled networks based on all relations. In particular, it is not possible to predict DTI based on AR only, because only the relationship between target and action is AR relationship. To verify the performance of our model on unlabeled networks, we conducted a third experiment, (3) the prediction performance of unlabeled networks based on all relations. To verify the robustness of our model, we performed a fourth experiment, (4) prediction performance based on the other datasets.

About evaluation metrics, we use AUC and AUPR to evaluate the performances of prediction.

5.2 Baseline Methods

DT-Hybrid [31] is a recommended method relying on network-based inference, which is based on domain knowledge, including drug and target similarity.

BLMNI [32] improves the traditional BLM method and can be used to deal with the new drug and target candidate problems, and it is called neighbor-based interaction-profile inferring.

HNM [33] combined with the drug target information, the intensity between the drug-disease pair is calculated by the iterative algorithm on the heterogeneous graph.

MSCMF [10] uses multiple drug similarity matrices and multiple target similarity matrices to project drugs and targets into a common low-dimensional feature space to predict DTI.

NetLapRLS [34] is a semi-supervised learning method—Laplacian regularized least square (LapRLS), which use Laplacian Regular Least Squares (LapRLS) to simultaneously use a small amount of available labeled data and a large amount of unlabeled data to obtain maximum generalization ability from the chemical structure and genome sequence.

DTINet [35] is a network integration approach that integrates heterogeneous information of drug-target heterogeneous networks.

RHINE [13] is a heterogeneous information network (HIN) embedding method which using the structural characteristics of heterogeneous relations.

NeoDTI [20] integrates diverse information from heterogeneous networks and use graph neural network to learn the representation of drugs and targets automatically.

EEG-DTI [36] propose an "end-to-end" learning framework based on heterogeneous graphical convolutional networks to learn low-dimensional feature representations of drugs and targets.

5.3 Task 1: Predictive Performance of Labeled Network Based on PR Relationships Only

After analyzing the dataset, as shown in Table 2, we found that only the type of target-action relationship is AR type, while most of the relationships in the drug-target heterogeneous network are PR type. And since our task is to predict the interaction relationship between drug and target protein, here we temporarily disregard the target-action relationship of AR type and only use the relationship of PR data, {drug-drug, drug-target, drug-disease, drug-side effect, target-target, target-disease} and compare the performance of our model with the other DTI prediction models.

During the experiment, we used 10% of the drug-target relationship and all other PR relationships as the training set, and the remaining 90% of the drug-target relationships was held out as the test set. According to the difference between positive and negative examples, we conducted two different experiments, the first one in which the ratio between positive and negative samples was set to 1:10, the other in which all unknown drug-target interacting pairs were considered as negative samples. Since the EEG-DTI model must consider all negative sample relationship pairs, experiments with AUC (1:10) are not supported.

The comparison results between our model and other models are shown in Table 3. The AUC scores obtained by our model in two different scenario prediction experiments are 94.3% and 95.8%, which exceeds the method NeoDTI by 3% and 2% respectively. Compared to NeoDTI, the embedding dimension of our method is 300, and NeoDTI is 1, 024.

What needs to be explained here is that in the NeoDTI experiment, in addition to the data mentioned above, the similarity information of the drug structure and the similarity information of the protein sequence are also used. Furthermore, NeoDTI is very time consuming and its running time is about 100 times that of our method.

Table 3. Performance evaluation of different models based on PR relations.

Method	AUC (1:10)	AUC (all)
MSCMF*	0.831	0.849
DT-Hybrid*	0.842	0.833
BLMNI*	0.855	0.850
HNM*	0.891	0.890
NetLapRLS*	0.905	0.895
DTINet*	0.919	0.909
NeoDTI*	0.941	0.913
EEG-DTI*	-	0.831
RTHNE_DTI*	**0.958**	**0.943**

5.4 Task 2: Predictive Performance of Labeled Network Based on All Relationships

In this task, we consider the target-action AR-type relationship, modeled individually for the characteristics of this relationship type, and incorporated it into the model of the PR-type relationship in task1. We

compare it with more advanced approaches. As before, we still use the 10% drug-target relationship and all other relationships as the training set, and the remaining data is used as the test set. For a fair comparison, we set the embedding dimension d = 100, because the two models run the most efficiently when the dimensionality is low, and all unknown pairs were targeted as the negative samples for all methods in this experiment. The results are shown in Table 4.

It can be seen from the results that our model is superior to the other two methods. Here, the NeoDTI method also utilizes the similarity information between the drug and the target protein, but the AUC value of our model is still 9% higher than it. Compared with EEG-DTI, our model fully considers the difference between AR relationship and PR relationship, and the AUC value is 11% higher than it. Compared with the RHINE method, our method considers more heterogeneous relationships. In terms of AUPR metrics, our model also far exceeds the RHINE model and is on par with the NeoDTI model. Thus, our results have better performance in this task. In addition, in this experiment, the AUC value of our method is 96.93%, which is about 3% higher than the result using only the PR relationship. It proves that the AR type relationship is also very important to improve the prediction ability.

Table 4. Performance evaluation of different models based on all relations.

Method	AUC	AUPR
NeoDTI*	0.883	0.288
EEG-DTI*	0.854	0.600
RHINE	0.923	0.145
RTHNE_DTI*	**0.958**	**0.264**

5.5 Task3: Predictive Performance of Unlabeled Network Based on All Relationships

In the existing DTI prediction methods, the drug and target pairs with known relationships are added to the training set to train the model. However, we assume whether it is possible to not add the relationship for prediction in the training set and only use others. In order to verify this conjecture, we conducted an experiment on task 3.

For task 3, We use all drug-target relationship pairs as test sets. The ratio of positive and negative samples is 1:10. Through experiments, our model's AUC and AUPR scores are 92.11% and 63.69%, respectively. Therefore, we can use their external relationships to predict when we have no clue whether there is an interaction between a drug and a target.

As shown in Figure 4, because we removed the target-drug interaction relationship in the task 3 experiment, the AUC value of our model RTHNE_DTI in task 3 (without label) is lower than that in task 2 (with label), but still much higher than the results of the NeoDTI model and RHINE model in task 2 (with label).

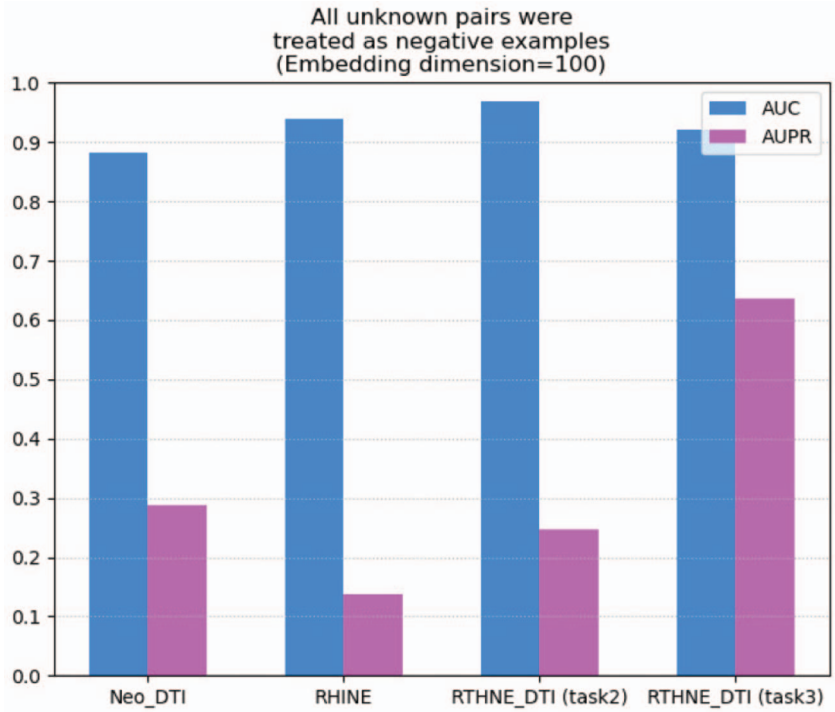


Figure 4. Performance evaluation of unlabeled network.

In addition, in terms of AUPR metrics, the results of our model RTHNE_DTI in task 3 (without labels) are the best, much higher than the results of all models in task 2 (with labels).

5.6 Predictive performance based on other datasets

DBLP is an integrated database system of computer English literature with the author as the core of the research results in the computer field. The details of the DBLP dataset are shown in Table 5.

Table 5. Statistical analysis of DBLP dataset.

Nodes	Num of Nodes	Relations	Num of Relations	Avg.t_u	Avg.t_v	D(e)	S(e)	Relation-ship type
Term(T)	8811	P-C	14376	1.0	718.8	718.8	0.05	AR
Paper(P)	14376	A-PC	24495	2.9	2089.7	720.6	0.08	AR
Author(A)	14475	P-P	41794	2.8	2.9	1.0	0.0002	PR
Conference(C)	20	D-Di	88683	6.2	10.7	1.7	0.007	PR
		P-A	260605	18.0	29.6	1.6	0.002	PR

From Table 5, we can see that the DBLP dataset contains more PR relations. We respectively predict the two relationship pairs author-author (A-A) and author-conference (A-C) in this experiment. The result is shown in Table 6.

Table 6. Performance evaluation of different datasets.

Dataset	AUC
DBLP(A-A)	0.924
DBLP(A-C)	0.906
Drug-target dataset (D-T)	0.969

The above experiments demonstrate that our method not only has good performance on drug networks, but also can achieve good results on scholar networks, and AUC value of our model do not fluctuate much on different datasets. It shows that our model has good robustness.

6. CONCLUSION AND FUTURE WORK

Accurately predicting the interaction between drugs and targets is important for drug research and development. In this paper, we apply the method of heterogeneous network representation learning to predict drug-target interactions. We build a heterogeneous network by the rich external relationships between drugs and target proteins and learn about drug and protein representations through neighboring nodes. We use data intelligence methods to divide the relationships into two categories: Affiliation relations and Peer relations, based on the different topologies of the relationships in the heterogeneous network, and model them separately. By doing this, our model can better capture the topological and semantic information of drug network in the same time, thus taking shorter time and achieving better results. Furthermore, the RTHNE_DTI model plays an important role in the real world. For example, we used RTHNE_DTI to discover a novel interaction between the drug Acemetacin and the target protein PTGS1, which has been proved to be correct in Drugbank database. It has proven to provide a powerful and useful tool for the drug discovery and drug repositioning process. In the future, we will consider the rich domain knowledge of drugs and proteins based on heterogeneous networks to further enhance the predictive effect of RTHNE_DTI and validate some of the predictions by wet lab experiments.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China, grant number 61402220, the key program of Scientific Research Fund of Hunan Provincial Education Department, grant number 19A439, the Project supported by the Natural Science Foundation of Hunan Province, China, grant number 2020JJ4525 and grant number 2022JJ30495.

AUTHOR CONTRIBUTION STATEMENT

Ll. Z (zhanglinlin@stu.usc.edu.cn, ORCID: 0000-0002-5035-3571) was jointly responsible for the investigation of the study, organization of data, design, analysis, and validation of methods, preparation of resources, and preparation of the original manuscript. Cp. Ouyang (ouyangcp@126.com, ORCID: 0000-0002-2154-0079) was jointly responsible for the investigation of the study, organization of the data, the

conceptualization of the methods, and review and editing of the manuscript, as well as supervision and project management. Fy. H (fyhu1124@163.com, ORCID: 0000-0002-1004-6913) contributed to the investigation of the study, the conceptualization of the methods, analysis, validation, and data curation. Yb. L participated in the model optimization and experimental design and analyzed the results. Z.G (gao27@indiana.edu) provided important feedback and edited the manuscript.

REFERENCES

[1] Kapetanovic, I.: Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. Chemico-Biological Interactions 171(2), 165–176 (2008)

[2] Pathak, J., Kiefer, R.C., Chute, C.G.: Mining drug-drug interaction patterns from linked data: A case study for warfarin, clopidogrel, and simvastatin. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 23–30 (2013)

[3] Ding, Y., Tang, J., Guo, F.: Identification of drug-target interactions via multiple information integration. Information Sciences 418, 546–560 (2017)

[4] D'Souza, S., Prema, K., Balaji, S.: Machine learning models for drug–target interactions: current knowledge and future directions. Drug Discovery Today 25(4),748–756 (2020)

[5] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug–target interaction net-works from the integration of chemical and genomic spaces. Bioinformatics 24(13), i232–i240, (2008)

[6] Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics 34(17), i821–i829 (2018)

[7] Wang, Y.-B., You, Z.-H., Yang, S., Yi, H.-C., Chen, Z.-H., Zheng, K.: A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. BMC Medical Informatics and Decision Making 20(2), 1–9 (2020)

[8] Chen, X., Yan, C.C., Zhang, X., Zhang, X., Dai, F., Yin, J., Zhang, Y.: Drug–target interaction prediction: databases, web servers and computational models. Briefings in Bioinformatics 17(4), 696–712 (2016)

[9] Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., Huang, L.F., Lewis, S.J., Nussinov, R., Cheng F.: Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. Bioinformatics 36(9), 2805–2812 (2020)

[10] Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1025–1033 (2013)

[11] Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H., Clevert, D.-A., Hochreiter, S.: Large-scale comparison of machine learning methods for drug target prediction on chembl. Chemical Science 9(24), 5441–5451(2018)

[12] Lee, I., Keum, J., Nam, H.: Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Computational Biology 15(6), e1007129, (2019)

[13] Lu, Y., Shi, C., Hu, L., Liu, Z.: Relation structure-aware heterogeneous information network embedding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4456–4463 (2019)

[14] Shang, J., Qu, M., Liu, J., Kaplan, L.M., Han, J., Peng, J.: Meta-path guided embedding for similarity search in large-scale heterogeneous information networks (2016)

[15] Fu, T.-Y., Lee, W.-C., Lei, Z.: Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1797–1806 (2017)

[16] Han, X., Shi, C., Wang, S., Philip, S.Y., Song, L.: Aspect-level deep collaborative filtering via heterogeneous information networks. In IJCAI, pp. 3393–3399 (2018)
[17] Muller, E., Peres, R.: The effect of social networks structure on innovation performance: A review and directions for research. International Journal of Research in Marketing 36(1), 3–19 (2019)
[18] Bu, Y., Huang, Y., Lu, W.: Loops in publication citation networks. Journal of Information Science 46(6), 837–848 (2020)
[19] Jin, S., Zeng, X., Xia, F., Huang, W., Liu, X.: Application of deep learning methods in biological networks. Briefings in Bioinformatics 22(2), 1902–1917 (2020)
[20] Wan, F., Hong, L., Xiao, A., Jiang, T., Zeng, J.: Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics 35(1), 104–111 (2019)
[21] Wasserman, S., Faust, K., et al.: Social network analysis: Methods and Applications 8, (1994)
[22] Faust, K.: Centrality in affiliation networks. Social Networks 19(2), 157–191 (1997)
[23] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, pp. 2787–2795 (2013)
[24] Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In 2012 IEEE 12th International Conference on Data Mining, pp. 1170–1175 (2012)
[25] Danielsson, P.-E.: Euclidean distance mapping. Computer Graphics and Image Processing 14(3), 227–248 (1980)
[26] Hsieh, C.-K., Yang, L., Cui, Y., Lin, T.-Y., Belongie, S., Estrin, D.: Collaborative metric learning. In Proceedings of the 26th international conference on world wide web, pp. 193–201 (2017)
[27] 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.: Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Research 46(D1), D1074–D1082 (2018)
[28] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, B.R., Shafreen, A. Venugopal: Human protein reference database—2009 update. Nucleic Acids Research 37(suppl 1), D767–D772 (2009)
[29] Davis, A.P., Murphy, C.G., Johnson, R., Lay, J.M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Rosenstein, M.C., Wiegers T.C.: The comparative toxicogenomics database: update 2013. Nucleicacids Research 41(D1), D1104–D1114 (2013)
[30] Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology 6(1), 343 (2010)
[31] Alaimo, S., Pulvirenti, A., Giugno, R., Ferro, A.: Drug–target interaction prediction through domain-tuned network-based inference. Bioinformatics 29(16), 2004–2008 (2013)
[32] Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., Zheng, J.: Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics 29(2), 238–245 (2013)
[33] Wang, W., Yang, S., Zhang, X., Li, J.: Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30(20), 2923–2930 (2014)
[34] Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In BMC Systems Biology, p. S6 (2010)
[35] Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., Zeng, J.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature Communications 8(1), 1–13 (2017)
[36] J.G., Peng, J., Wang, Y.: An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. Briefings in Bioinformatics 22(5), 1–9 (2021)

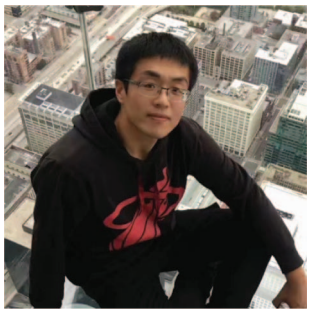
AUTHOR BIOGRAPHY

Linlin Zhang received her B.S. degree in Software Engineering from the School of Computer Science, University of South China, China, in 2020. She is pursuing her M.S. degree with a specialization in software engineering at the University of South China. Her research interests include network representation learning and drug discovery.
ORCID: 0000-0002-5035-3571

Chunping Ouyang received a Ph.D. degree from the University of Science & Technology Beijing, China, in 2011. From 2017 to 2018, she was a visiting scholar at Indiana University Bloomington. She is a professor of computer science at the University of South China and supervisor of postgraduate. Her main research interests are semantic web technology, knowledge graphs, natural language processing, and domain data analysis.
ORCID: 0000-0002-2154-0079

Fuyu Hu received her master's degree in computer technology from the School of Computer Science, South China University, China, in 2021. Her research interests include network representation learning and drug discovery. She is working on a related project at the Bank of Communications in China.
ORCID: 0000-0002-1004-6913

Yongbin Liu received a Ph.D. degree from the University of Science & Technology Beijing, China, in 2013. From 2013 to 2015, he was a post-doc research fellow at Tsinghua University. He is an associate professor at the University of South China. His research interests include natural language processing and knowledge engineering.
ORCID: 0000-0002-3369-3101



Zheng Gao is an Applied Scientist at Amazon Alexa AI. He received his Ph.D. degree in Information Science and minor in Computer Science from Indiana University Bloomington, advised by Prof. Xiaozhong Liu in 2020. His research interests are primarily in the area of Graph Mining and Natural Language Processing (NLP). Particularly, he is applying deep learning techniques on the interdisciplinary field therein them to solve Community Detection, Information Retrieval and Recommendation related tasks. He is currently working as an applied scientist at Amazon Alexa AI and build NLU models to handle customer utterances.
ORCID: 0000-0001-7549-033X