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

Fuyu Hu
School of Computer，University of South China
Chunping Ouyang（ ouyangcp@126.com ）
School of Computer，University of South China
Yongbin Liu
School of Computer，University of South China
Zheng Gao
Luddy School of Informatics, Computing, and Engineering，Indiana University
Yaping Wan
School of Computer，University of South China

Research Article

Keywords: Link prediction, Heterogeneous information network, Drug-target interaction

Posted Date: January 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-147527/v1

RESEARCH

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

Fuyu Hu[1], Chunping Ouyang[1,2*], Yongbin Liu[1,2], Zheng Gao[3] and Yaping Wan[1,2]

[*]Correspondence:
ouyangcp@126.com
[1]School of Computer, University of South China, Hengyang,Hunan, 421001, China
[2]Hunan provincial base for scientific and technological, Innovation Cooperation, Hengyang,Hunan, 421001, China
Full list of author information is available at the end of the article

Abstract

Background: Predicting interactions between drugs and target proteins is a key task in drug discovery. Although the method of validation via wet-lab experiments has become available, experimental methods for drug-target interactions (DTIs) 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. Usually, we construct a heterogeneous network with drugs and target proteins to calculate the relationship between them. However, most calculation methods do not consider the topological structure of the relationship between drugs and target proteins. Fortunately, Network Embedding Learning provides new and powerful graph analytical approaches for predicting drug-target interaction, which is considering both content and topology of network.

Results: In this article, we propose a relational topology-based heterogeneous network embedding method to predict DITs, abbreviated as RTHNE_DTI. We use the ideas of word embeddings to turn heterogeneous network with drugs and target proteins into dense, low-dimensional real-valued vectors. Furthermore, according to two different topological structure of the relationship between the nodes, we represent them separately by training two different models.Then the meaningful vectors represented for drugs and target proteins can be used to calculate the interaction of them easily. Results show that by considering topological structure and different relationship type of drugs and target proteins, RTHNE_DTI outperforms other state-of-the-art methods on both labeled network and unlabeled network.

Conclusions: This work proposes heterogeneous network representation learning for DITs prediction. To the best of our knowledge, this study first introduces relation classification to heterogeneous network embedding to improve predicting DTIs efficiently.

Keywords: Link prediction; Heterogeneous information network; Drug-target interaction

Introduction

The prediction of drug-target interactions (DTIs) is the key to the development of new drugs. It plays an important role in the study of drug toxicity and side effects and in the treatment of diseases. However, traditional methods based on large-scale biological experiments usually take several years and are often very expensive [1]. In recent years, with the rapid development of computer technology and the accumulation of large amounts of medical 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 similarity-based [5], deep learning based [6, 7], and network-based [8, 9] methods.

In the branch of similarity-based methods, Yamanashi et al. [5] proposed a method strategy mainly using nuclear regression methods, taking the information of known drug interactions as input new DTIs. This is a method proposed earlier to reveal the significant correlation between drug structure similarity, target sequence similarity, and drug-target interaction network topology. Zheng et al. [10] proposed a method to predict the interaction between drugs and targets by using a multi-similarity cooperation matrix. The core idea is to get the interaction matrix and similarity score matrix by multiplying two similarity matrices representing drugs and targets. As another similarity-based approach to DTIs prediction, compared with the previous method, Ezzat et al. [11] considered that many edges not appearing in the network are actually unknown or missing, the method of graph regularization matrix decomposition is used to predict unknown edges. However, most of these methods employed the chemical structure and protein sequence of the drug. However, in public data sets, it is often difficult to obtain the protein sequence and chemical information of many polymers.

Recently, the rapid development of deep learning technology has offered effective ways of predicting DTIs. Mayr et al. [12] 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. [13] 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.

There are various networks in practice, such as social networks [14], citation networks [15], and biological information networks [16]. And some interesting research works on network analysis have attracted increasing attention. Particularly, link prediction is one of the hotspot tasks of network analysis. Currently, most network-based DTIs prediction is based on machine learning [8]. Wang et al. transformed new DTIs prediction problems into a two-layer graphical model named the restricted Boltzmann machine (RBM). Wan et al. [17] developed a new nonlinear end-to-end learning model, called Neo_DTI, which integrates different heterogeneous information of drugs and targets, and learned the representation of drugs and targets to predict DTIs. However, note that one of the drawbacks of these methods is that they may not work when the chemicals pathway and proteins interact is unknown.

Heterogeneous network representation learning is a hot topic in current research and has quite good performance in link prediction [18]. Although heterogeneous network representation learning methods have been widely adopted for link prediction of social networks with good results, they have not been used for link prediction of DTIs to be best of our knowledge. Most of the previous studies on networks were based on homogeneous networks, to be specific, nodes in networks are of the same type. With the development of network representation learning, to simulate the

heterogeneity of networks, some people have tried heterogeneous network representation learning. For instance, Shang et al. [19] proposed a framework, namely ESim, which uses random walks based on the mate-path to generate node sequences to optimize the similarity between several points. Fu et al. [20] proposed a method of heterogeneous information network representation HIN2vec, unlike many previous works based on the skip-gram language model, and the core of HIN2Vec is a neural network model, which learns the representation of both nodes and relationships (meta-paths) in a network. Han et al. [21] proposed an aspect-level collaborative filtering model based on neural networks. In their model, they extracted similarity matrices of different aspect levels of nodes through different meta-paths, and inputted these matrices into the designed deep neural network to learn the potential factors of the aspect level. Usually these methods are used in social networks, scholar networks, etc.. Due to the drug network is very similar to these networks in structure, so we try to solve the problem of drug-target prediction in this way. Particularly, our main contributions are as follows:

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

• The traditional heterogeneous network representation learning method uses a single model to deal with all relationships. Here according to the different topological structure of the relationship between nodes in a heterogeneous network, we divide the relationship into two types: Affiliation relationship and Peer relationship, and built different models for them to capture the rich semantic information between nodes better. And make full use of the diversity of drug network relationships.

• 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 extensive experiments using real drug data set and compare with other predictive models and the results show that RTHNE_DTI has the best predictive performance.

1 Problem Formulation

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

Defintion 1: Heterogeneous Network (HN).

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

Defintion2: 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:

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

Defintion3: 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 $d \ll \mid V \mid$ is the dimension of the representation space.

2 Method

2.1 Overall Framework

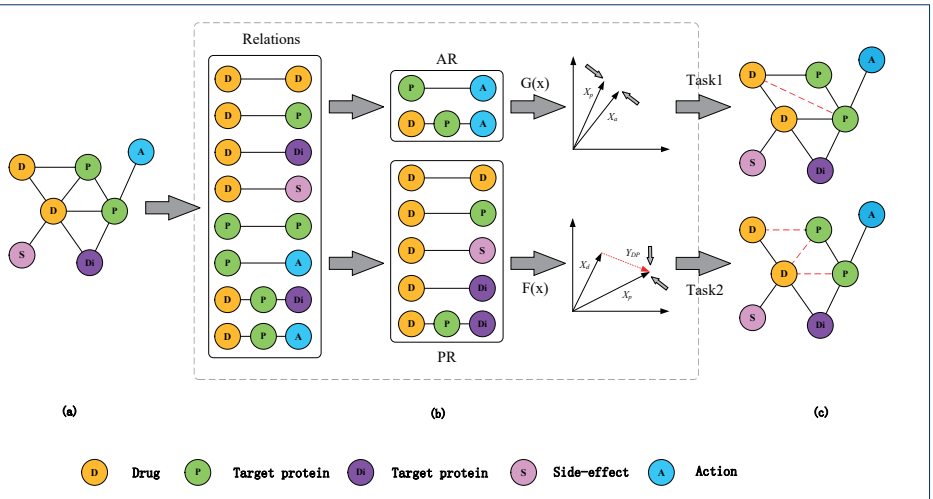

Figure 1 The overall framework

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.

2.2 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 equal number of nodes of the two types of connections, and some relationship pairs had significantly different number of nodes of the two types of connections, as shown in the 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 very large, so 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.

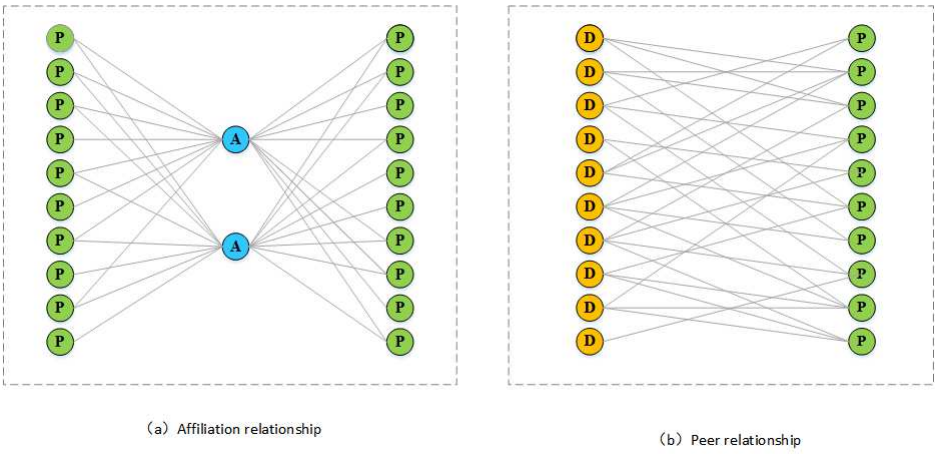

Figure 2 Different relationship topology

The two types of nodes do not differ greatly in number, so they form a well-balanced network. As shown in Figure 2 (b).

To make full use of 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 [22]. 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:

$$D(\mathrm{e}) = \frac{\max(\overline{d}_{n_u}, \overline{d}_{n_v})}{\min(\overline{d}_{n_u}, \overline{d}_{n_v})} \tag{2}$$

Where n_u, n_v represent the node type of nodes u, v in a retion 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 $D(r)$ value show stronger affiliation relationship (AR) between them, and nodes connected by this relationship have more similar properties [23]. 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:

$$S(\mathrm{e}) = \frac{N_{\mathrm{r}}}{N_{\mathrm{n}_u} \times N_{n_v}} \tag{3}$$

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

Table 1 Statistical analysis of dataset

Nodes	Number of Nodes	Relations	Number 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

emphasized that in this way, these relationships can also be consistently divided into two categories, PR and AR.

We conducted a comprehensive analysis of the obtained data according to the above indicators as shown in Table 1.

2.3 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 [24]. We use the Euclidean distance method for the AR mainly because of the following reasons. First, the nodes connected by this relationship share the same attributes [25], so the nodes connected by the AR can be directly approached in the vector space, which is consistent with the Euclidean distance optimization [26]. 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 [27], 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 1, we found that in the drug heterogeneous network we 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:

$$f(\mathrm{a}, b) = w_{a,b}||P_a + \mathrm{Q}_r - P_b|| \tag{4}$$

For the relationship tuples $(a, r, b) \in T_{PR}$ whose relationship is PR in the heterogeneous network, the marginbased loss function [24] is defined as follows:

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

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. For the heterogeneous network we constructed, only the target protein and its action type belong to the AR relationship. 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:

$$\mathrm{g}(m,n) = w_{m,n} ||P_m - P_n||_2^2 \tag{6}$$

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:

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

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

2.4 Conjunctive Model

To make the model more complete, we smoothly join the two models presented in the previous section by using a loss-minimization function.

$$L = L_{PR} \oplus L_{AR} \tag{8}$$

In Table 1, we can clearly find that the distribution of PR and AR are quite unbalanced. In the second experiment, to prevent the traditional edge sampling method from biasing to the larger number, we use probability distribution to extract positive samples. We adopt the previous method to construct a negative sample, the form is as follows:

$$T'_{(m,n)} = \{(m', i, n) | m' \in N\} \cup \{(m, i, n') | n' \in N\} \tag{9}$$

The above formula shows that for the positive relationship tuple (m, i, n), m, n can be randomly replaced, but not simultaneously.

3 Experiments

3.1 Datasets

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. [28].

HPDR (Human Protein Reference Database) contains manually curated scientific information pertaining to the biology of most human proteins, and the data of protein?protein interactions extracted from the HPRD database Release 9 [29].

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 [30].

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 [31].

We obtained data from the above four sources, and after data preprocessing, we finally obtained 708 drugs, 1512 target proteins, 5603 diseases, 47 actions, and 4192 side effects. Some descriptive statistics of the dataset are shown in Table 1.

3.2 Experimental Settings

RTHNE_DTI has three parameters: embedding dimension d, the margin γ, and α, we set γ=1, and α=0.01. To study the influence of different dimensions on our model, we explored the 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.

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

3.3 Baseline Methods

DT-Hybrid [32] is a recommendation method relying on network-based inference, which is based on domain knowledge including drug and target similarity.

BLMNI [33] improves the traditional BLM method and can be used to deal with new drug and target candidate problems, and it called neighbor-based interaction-profile inferring.

HNM [34] 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 DTIs.

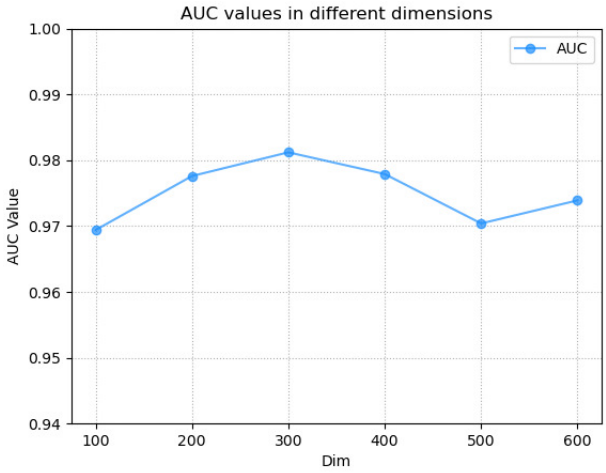

Figure 3 Parameter Analysis.

NetLapRLS [35] 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 chemical structure and genome sequence.

DTINet [36] is a network integration approach that integrates heterogeneous information of drug-target heterogeneous networks.

RHINE [18] is a heterogeneous information network (HIN) embedding method which using the structural characteristics of heterogeneous relations.

Neo_DTI [17] integrates diverse information from heterogeneous networks, and use graph neural network to automatically learn the representation of drugs and targets.

3.4 Task 1 : Predictive performance of labled network based on PR relationships

In this paper, we conducted experiments in two different scenarios, one based on the labeled network and the other based on the unlabeled network to predict the interactions between drugs and targets.

After our analysis of the data set, as shown in Table 1, we found that most of the relationships in the drug-target heterogeneous network are PR-type, so here we only used the data with the relationship PRs specifically : {drug-drug, drug-target, drug-diseases, drug-side-effects, target-target, target-diseases}, and we compared the performance of our model with other DTIs 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 sample.

The comparison results between our model and other models are shown in Table 2. The AUC scores obtained by our model in two different scenario prediction experiments are 94.3% and 95.8%, which exceeds the method Neo_DTI by 3% and 2% respectively. Compared to Neo_DTI, the embedding dimension of our method is 300, and Neo_DTI is 1024.

What needs to be explained here is that in the Neo_DTI 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, Neo_DTI is very time consuming and its running time is about 100 times that of our method.

Table 2 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
Neo_DTI	0.941	0.913
RTHNE_DTI	**0.958**	**0.943**

3.5 Task 2 : Predictive performance of labled network based on all relationships

In this part we used all relations and compare with more advanced methods. As before, we still use the 90% 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 negative sample for all method in this experiment. The results are shown in Figure 4.

It can be seen from the results that our model is superior to the other two methods. Here, the Neo_DTI 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 the RHINE method, our method considers more heterogeneous relationships. 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.

3.6 Task3 : Predictive performance of unlabeled network based on all relationships

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

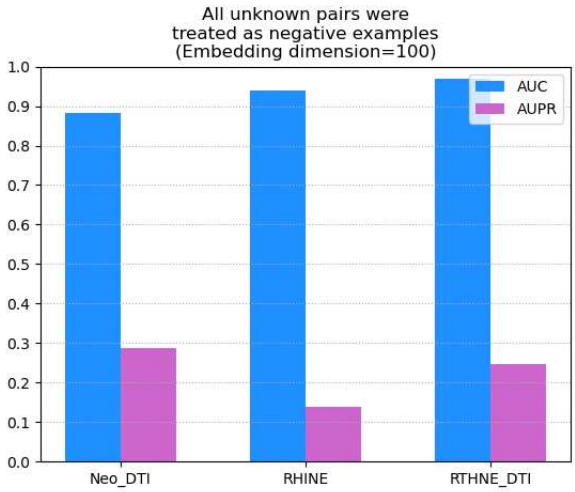

Figure 4 Performance evaluation of different models based on all relations.

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, the AUC and AUPR score obtained by our model are 92.11% and 63.69%, respectively. Therefore, we can use their external relationships to predict when we have no clue on whether there is an interaction between a drug and a target.

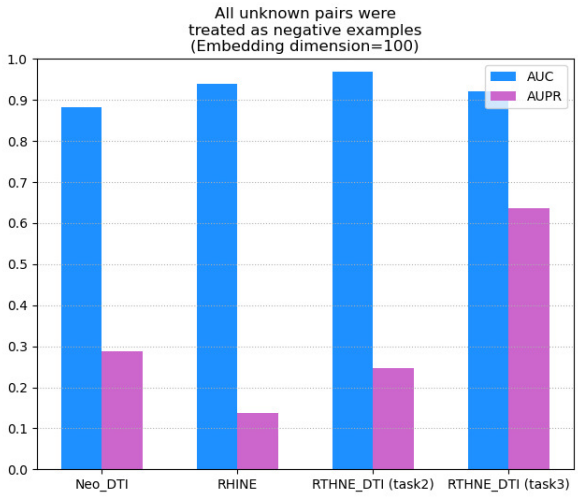

Figure 5 Performance evaluation of unlabeled network

As shown in Figure 5, although the AUC value of our method in task 3 is lower than that in task 2, it is very close to the result of Neo_DTI method. In addition, from the AUPR indicator, the result of RTHNE_DTI in task 3 is the best.

3.7 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 is shown in Table 3.

Table 3 Statistical analysis of DBLP dataset

Nodes	Number of Nodes	Relations	Number of Relations	Avg.t_u	Avg.t_v	$D(r)$	$S(r)$	Relationship 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	IR
Conference(C)	20	D-Di	88683	6.2	10.7	1.7	0.007	IR
-	-	P-A	260605	18.0	29.6	1.6	0.002	IR

From Table 3 we can see that DBLP dataset contains more IR relations. In the experiment, we respectively predict the two relationship pairs author-author, author-conference. The result as show in Table 4.

Table 4 Performance evaluation of different datasets

Dataset	AUC
DBLP(A-A)	0.924
DBLP(A-C)	0.906
RTHNE_DTI(D-T)	0.969

The above experiments proved that our method can not only achieve good results on the drug network, but also on the scholar network. Compared with the traditional drug-target prediction method, our method is more general.

4 Conclusions

Accurately predicting the interaction between drugs and target proteins is very important for drug discovery. In this paper, we applied the method of heterogeneous network representation learning to predict drug-target interactions. We built a heterogeneous network by the rich external relationships between drugs and target proteins, and learn about drug and protein representations through neighboring nodes. According to the different topological structures of the relationship in the heterogeneous network, we divided the relationship into two categories: Affiliation relations and Peer relations, and model them separately. By doing this, our model can better capture the topology information and semantic information of the network with drugs and target proteins.To evaluate the ability of our method, we compared it with several state-of-the-art approaches. The results proved that RTHNE_DTI taked shorter time and was more efficient on the DTIs task. Furthermore, whether it is labeled network or unlabeled network with the relationship of drugs and target proteins, RTHNE_DTI has achieved good results. In the future, we will consider the rich domain knowledge of drugs and target proteins on the basis of heterogeneous network to enhance the prediction effect.

Acknowledgements
The authors would like to thank all anonymous reviewers for their advice.

Funding
This work was supported by the National Natural Science Foundation of China, under Grants 61402220; the Key Program of Research Foundation of Education Bureau of Hunan Province, China, under Grants 19A439; the Natural Science Foundation of Hunan Province, China, under Grants 2020JJ4525; the Science-Technology Foundation of Hunan Province, China, under Grants 2020SK3010.

Abbreviations
DTIs:drug-target interactions; RTHNE:relational topology-based heterogenous network embedding; AR: the affiliation relationship; PR: the peer relationship

Availability of data and materials
The data and code is available at: https://github.com/Hufuyu1112/11111.git

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
Conceptualization, C.O., Y.L. and Z.G.; methodology, C.O. and F.H.; validation, F.H. and Y.L.; formal analysis, F.H. and Y.L.; investigation, C.O., F.H. and Y.L.; resources, Y.W.; data curation, C.O., F.H. and Y.L.; writing—original draft preparation, F.H.; writing—review and editing, C.O. and Y.B.; supervision, C.O.; and project administration, Y.W. All authors have read and agreed to the published version of the manuscript.

Author details
[1]School of Computer, University of South China, Hengyang,Hunan, 421001, China. [2]Hunan provincial base for scientific and technological, Innovation Cooperation, Hengyang,Hunan, 421001, China. [3]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA.

References

1. Kapetanovic, I.: Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. Chemico-biological interactions **171**(2), 165–176 (2008). doi:10.1016/j.cbi.2006.12.006
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). IEEE
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). doi:10.1016/j.drudis.2020.03.003
5. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13), 232–240 (2008). doi:10.1093/bioinformatics/btn162
6. Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics **34**(17), 821–829 (2018). doi:10.1093/bioinformatics/bty593
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). doi:10.1093/bioinformatics/btaa010
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. Ezzat, A., Zhao, P., Wu, M., Li, X.-L., Kwoh, C.-K.: Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM transactions on computational biology and bioinformatics **14**(3), 646–656 (2017). doi:TCBB.2016.2530062
12. 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). doi:10.1039/C8SC00148K
13. 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), 1007129 (2019). doi:10.1371/journal.pcbi.1007129
14. 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). doi:10.1016/j.ijresmar.2018.05.003
15. Bu, Y., Huang, Y., Lu, W.: Loops in publication citation networks. Journal of Information Science, 0165551519871826 (2019)
16. Jin, S., Zeng, X., Xia, F., Huang, W., Liu, X.: Application of deep learning methods in biological networks. Briefings in Bioinformatics (2020)
17. 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). doi:10.1101/261396
18. Lu, Y., Shi, C., Hu, L., Liu, Z.: Relation structure-aware heterogeneous information network embedding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4456–4463 (2019)

19. 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. arXiv preprint arXiv:1610.09769 (2016). doi:10.1145/1235
20. 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)
21. 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)
22. Wasserman, S., Faust, K., *et al.*: Social Network Analysis: Methods and Applications vol. 8. Cambridge university press, ??? (1994)
23. Faust, K.: Centrality in affiliation networks. Social networks **19**(2), 157–191 (1997)
24. 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)
25. 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). IEEE
26. Danielsson, P.-E.: Euclidean distance mapping. Computer Graphics and image processing **14**(3), 227–248 (1980)
27. 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). doi:10.1145/3038912.3052639
28. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.*: Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic acids research **46**(D1), 1074–1082 (2018). doi:10.1093/nar/gkx1037
29. Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.*: Human protein reference database—2009 update. Nucleic acids research **37**(suppl_1), 767–772 (2009). doi:10.1093/nar/gkn892
30. 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., *et al.*: The comparative toxicogenomics database: update 2013. Nucleic acids research **41**(D1), 1104–1114 (2013). doi:10.1093/nar/gks994
31. 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)
32. Alaimo, S., Pulvirenti, A., Giugno, R., Ferro, A.: Drug–target interaction prediction through domain-tuned network-based inference. Bioinformatics **29**(16), 2004–2008 (2013). doi:10.1093/bioinformatics/btt307
33. 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). doi:10.1093/bioinformatics/bts670
34. 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). doi:10.1093/bioinformatics/btu403
35. Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In: BMC Systems Biology, vol. 4, p. 6 (2010). Springer
36. 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). doi:10.1038/s41467-017-00680-8

Figures

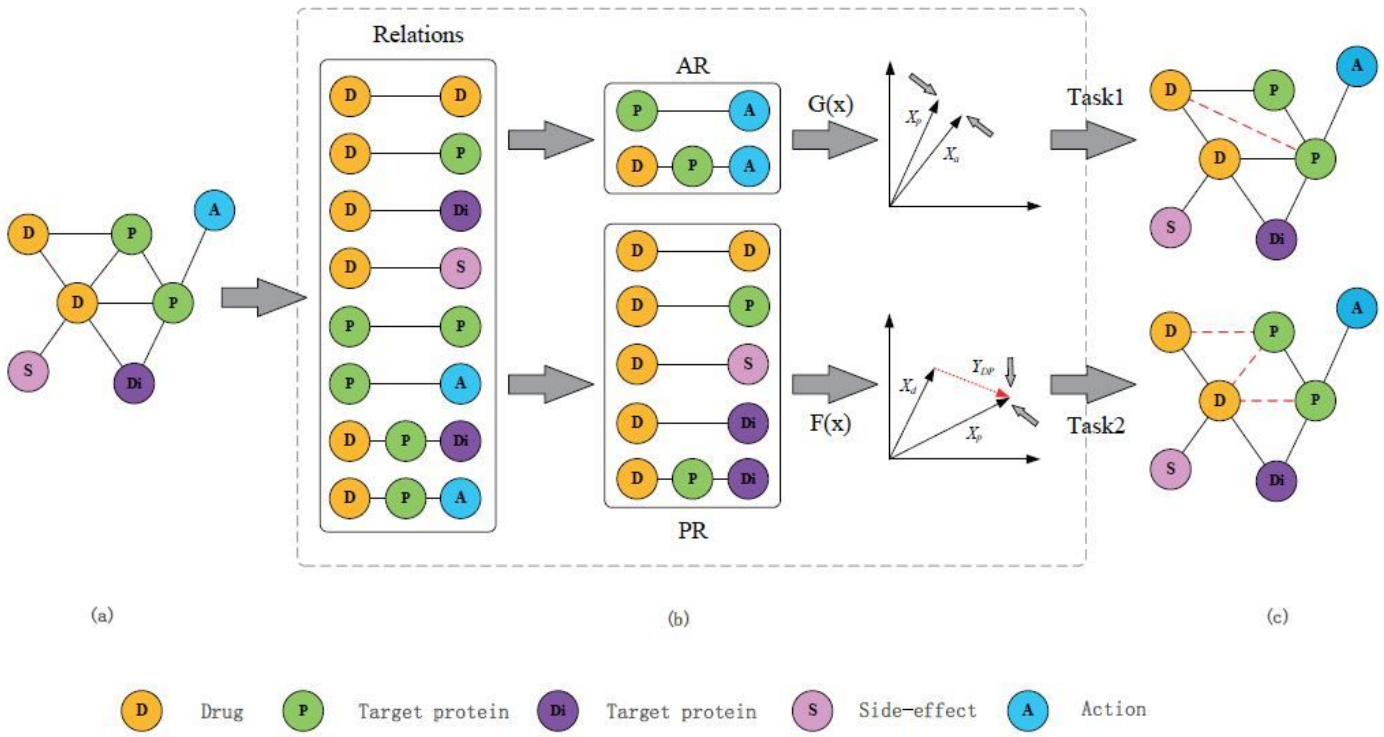

Figure 1

The overall framework

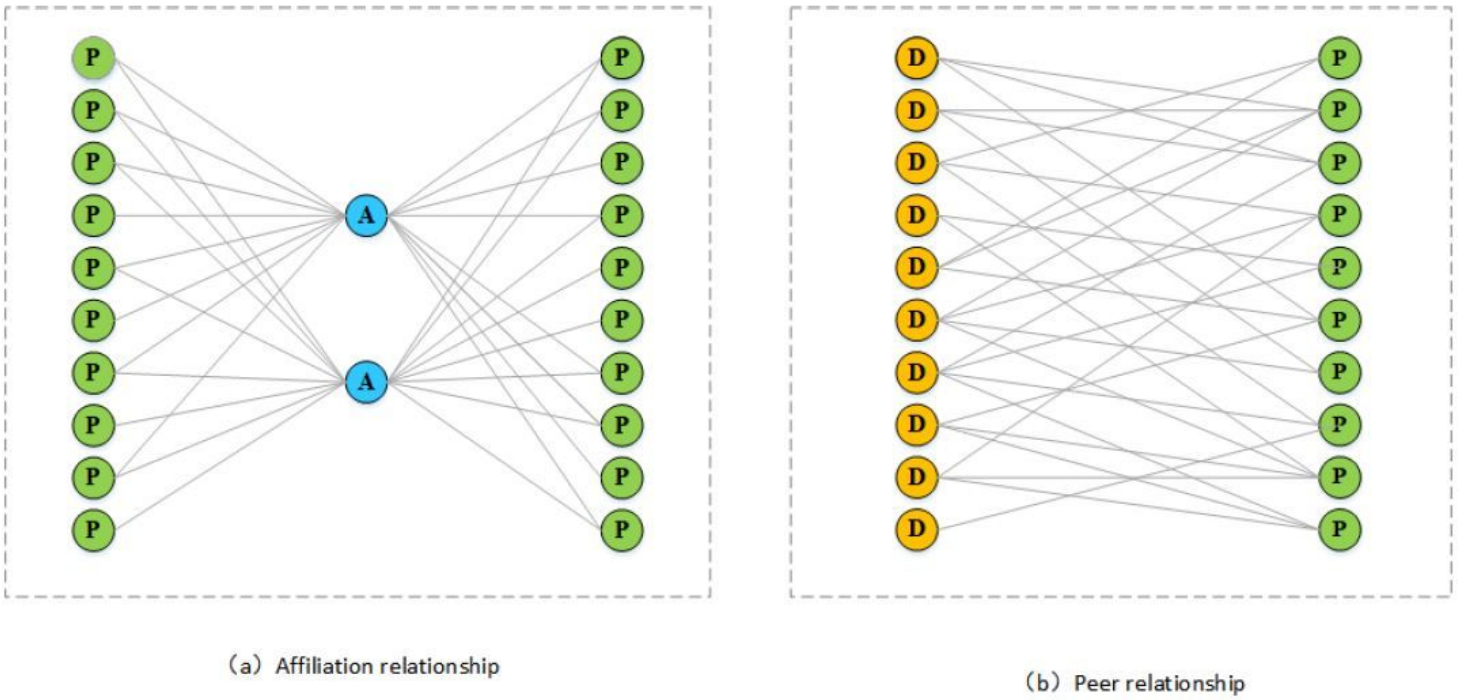

Figure 2

Different relationship topology

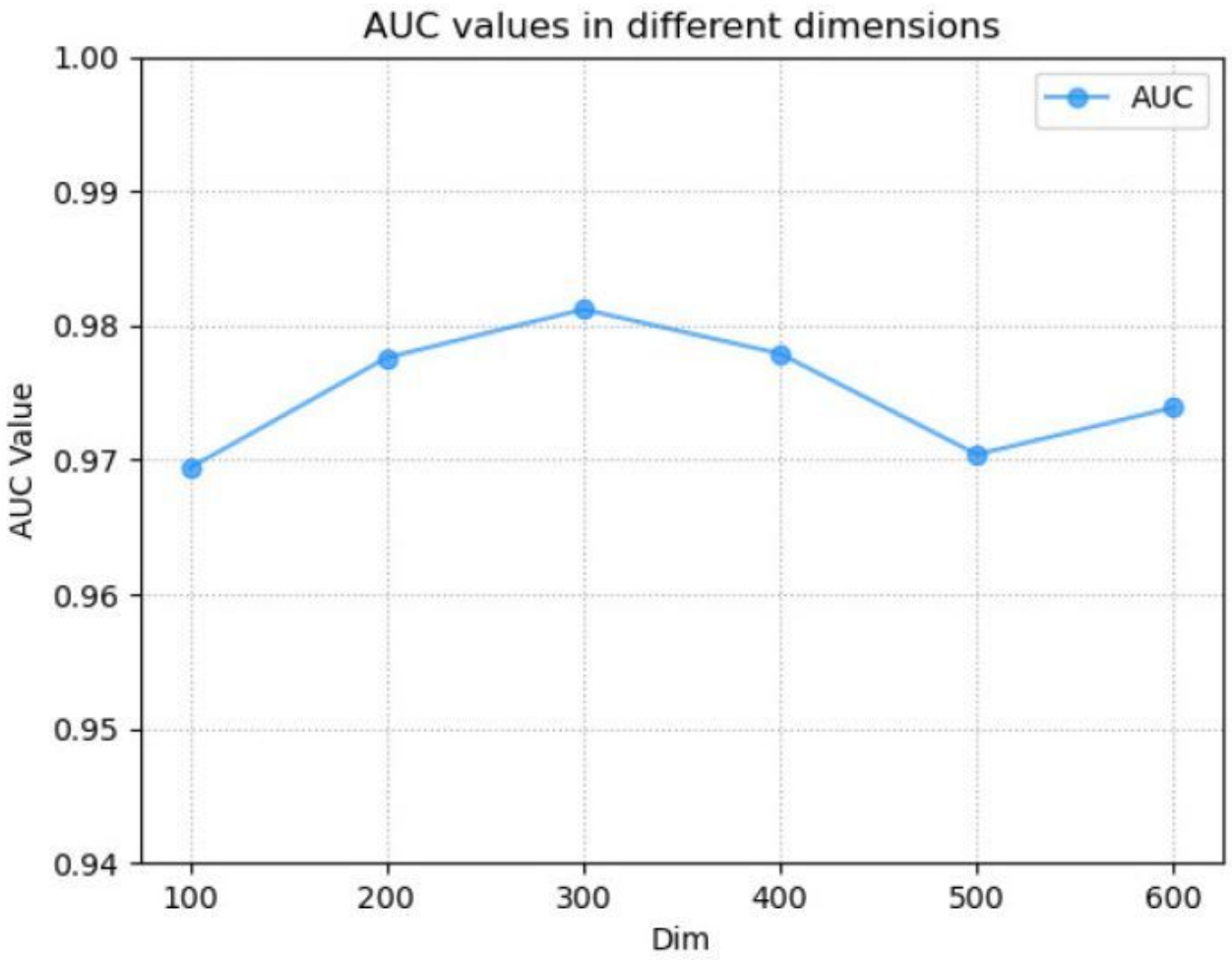

Figure 3

Parameter Analysis.

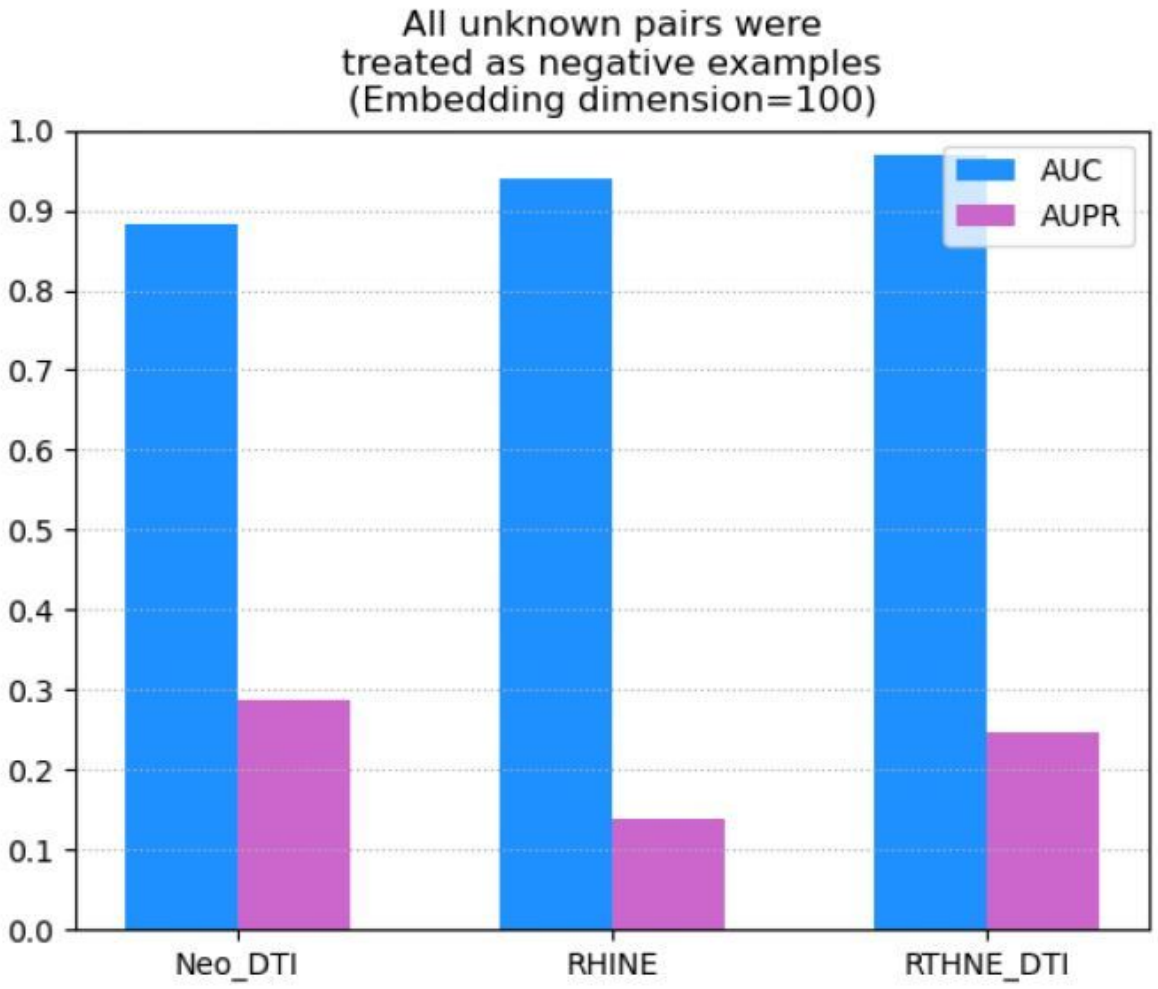

Figure 4

Performance evaluation of different models based on all relations.

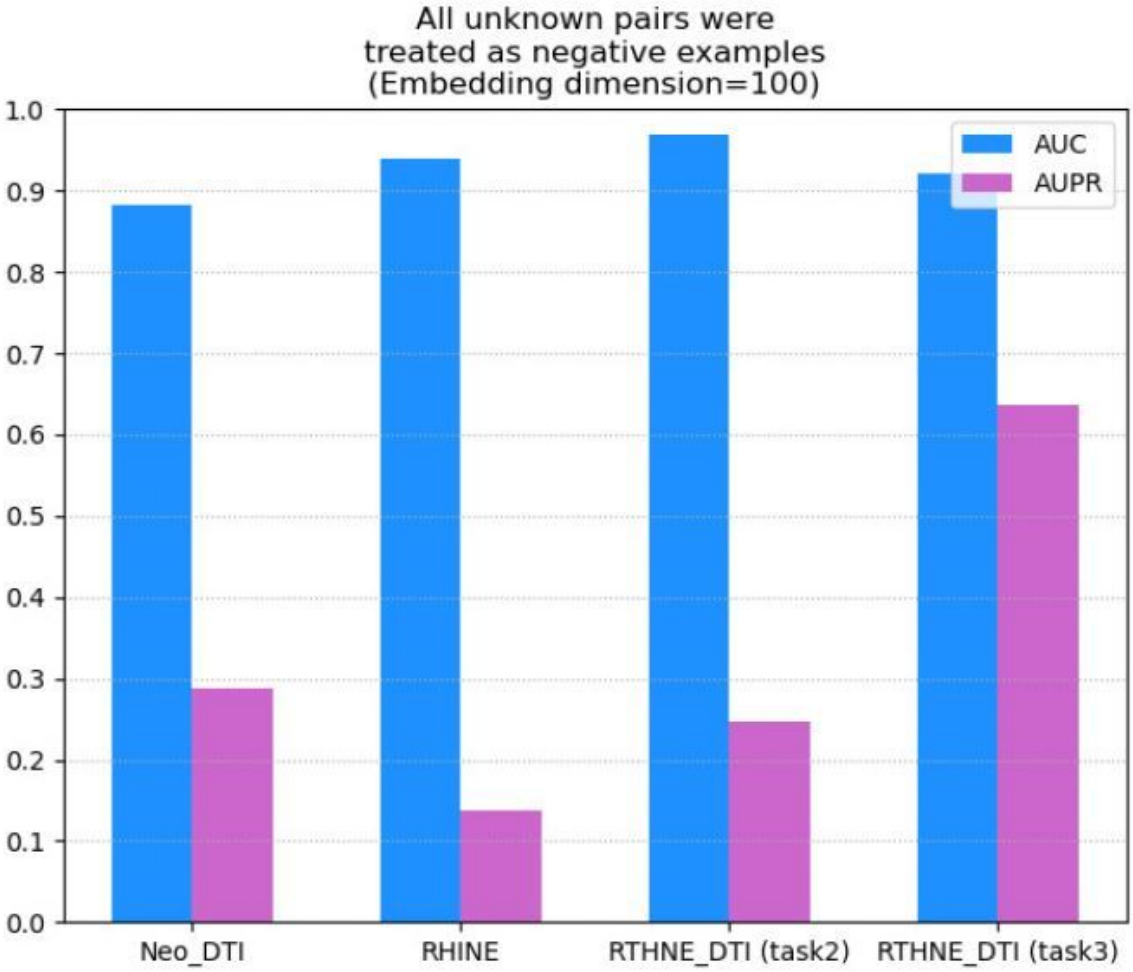

Figure 5

Performance evaluation of unlabeled network