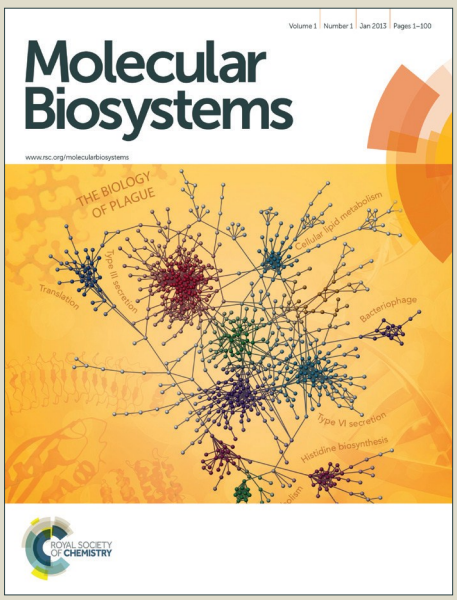
Molecular BioSystems

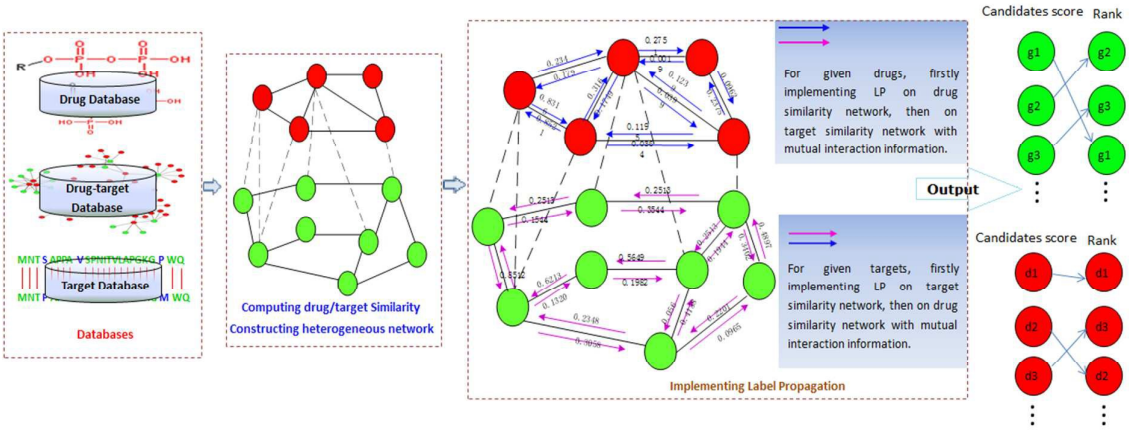
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

• A table of contents entry: graphic maximum size 8 cm x 4 cm and one sentence of text, maximum 20 words, highlighting the novelty of the work



By implementing Label Propagation on drug/target similarity network with mutual interaction information derived from drug-target heterogeneous network, LPMIHN algorithm identifies potential drug-target interactions.

Prediction of drug-target interaction by label propagation with mutual interaction information derived from heterogeneous network

Xiao-Ying Yan[1,2], Shao-Wu Zhang[1]*,Song-Yao Zhang[1]

1 Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China

2 College of Computer Science, Xi'an Shiyou University, Xi'an, 710065, China

* To whom correspondence should be addressed. Email: zhangsw@nwpu.edu.cn; Tel: +86-29-88431308

Abstract: Identification of potential drug-target interaction pairs is very important, which is not only for providing greater understanding of protein function, but also for enhancing drug research, especially for drug function repositioning. Recently, numerous machine learning-based algorithms (e.g. kernel-based, matrix factorization-based and network-based inference methods) have been developed for predicting drug-target interactions. All these methods implicitly utilize the assumption that similar drugs tend to target similar proteins, and yield better results for predicting interactions between drugs and target proteins. To further improve the accuracy of prediction, a new method of network-based label propagation with mutual interaction information derived from heterogeneous network, namely LPMIHN, is proposed to infer the potential drug-target interactions. LPMIHN separately performs label propagation on drug and target similarity networks, but the initial label information of target (or drug) network comes from drug (or target) label network and known drug-target interaction bipartite network. The independent label propagation on each similarity network explores the cluster structure in its network, and the label information from the other network is used to capture mutual interactions (bicluster structures) between the nodes in each pair of the similarity networks. Comparison with other recent state-of-the-art methods on the four popular benchmark datasets of binary drug-target interactions and two quantitative kinase bioactivity datasets, LPMIHN achieves the best results in terms of AUC and AUPR. In addition, many of the promising drug-target pairs predicted from LPMIHN are also confirmed on the latest

publicly available drug-target databases such as ChEMBL, KEGG, SuperTarget and Drugbank. These results demonstrate the effectiveness of our LPMIHN method, indicating that LPMIHN has the great potential for predicting drug-target interactions.

1 Introduction

Identification of drug-target interactions (DTIs) and compound-protein interactions (CPIs) is very important in drug research, which is helpful for discovering new drugs or identifying novel targets for existing drugs, and also may help understanding the causes of side effects of existing drugs. However, known drug-target interactions are very limited[1-3], e.g., one of the largest chemical compound database PubChem[4] contains around 35 million compounds, but only less than 7,000 compounds have target protein information, and existing databases such as ChEMBL[5], KEGG DRUG[6], SuperTarget[7] and DrugBank[8] include a small number of drug-target interaction pairs validated with experimental methods.

Generally, there are two ways used to find the drug-target interactions, one is biochemical experimental (or *in vitro*) methods, another is computational (or *in silico*) methods. However, experimental methods to determine drug-target interactions are usually time-consuming, tedious and expensive, and sometimes lack reproducibility [9-11]. Thus, it is highly desired to develop computational methods for efficiently and effectively analyzing and detecting new drug-target interaction pairs, which lead to appearing a variety of theoretical and computational methods in recent few years [12-39]. Computational methods can also guide experimentalists designing the best experimental scheme, narrowing the scope of candidate targets to accelerate drug discovery, and provide supporting evidence for their experimental results.

Molecular docking simulation and machine learning are the two major computational methods for predicting drug-target interactions. Docking simulation is widely accepted in biology, but it is not only heavily time-consuming, and also needs to know three-dimensional (3D) structures of targets[40], while many of them are still unavailable[14, 23], especially for membrane proteins. In contrast with docking

simulation, machine learning is much more efficient, which allows larger-scale predictions and tests a larger number of promising candidates of targets and drugs.

Machine learning-based methods developed for predicting drug-target interactions so far can be roughly classified into two types: feature vector-based machine learning and similarity-based machine learning[23, 24]. Feature vector-based machine learning methods use a vector of descriptors to represent each drug-target pair/non-interaction pair, and adopt one classifier (e.g. SVM, KNN) to predict the drug-target interactions [21, 22, 25-27, 30]. However, the performance of these methods severely depends on the generation of negative samples (i.e. non-interacting drug-target pairs) [21, 23]. Randomly generated negative samples may include real positive samples not yet unknown[21].

Similarity-based machine learning methods can be grouped into three categories such as kernel-based approaches [28, 29, 34, 41], matrix factorization-based approaches[31, 34] and network-based inference[32, 42-44]. Kernel-based approaches, such as net Laplacian regularized least squares (NetLapRLS) [29], regularized least squares with Kronecker product kernel (RLS-Kron)[28], pair kernel method (PKM)[41], kernel-based data fusion[36] and kernelized Bayesian matrix factorization with twin kernels (KBMF2K)[34], use the drug similarity information and target similarity information to construct kernels for predicting the drug-target interactions. Matrix factorization-based approaches, such as KBMF2K[34], multiple similarity collaborative matrix factorization (MSCMF)[31], project the drugs and targets into a common low-rank feature space by matrix factorization method for predicting the drug-target interactions.

Network-based inference methods, such as bipartite local model (BLM)[37], BLM with neighbor-based interaction profile inferring (BLM-NII)[32], domain tuned-hybrid (DT-Hybrid)[45], network-based inference(NBI)[35], weight network-based inference (WNBI)[46], network consistency-based prediction method (NetCBP)[43] and network-based random walk with restart on the heterogeneous network (NRWRH) [42], use graph and network theory to infer the drug-target interactions by constructing drug-target bipartite graph, drug similarity network and

target similarity network. BLM[37] transformed the edge-prediction problems into well-known binary classification problems, and combined the results from drug-based prediction and target-based prediction to obtain the final results. However, BLM is not able to provide a reasonable prediction results for drug/target candidates that are currently new, because a new drug (or target) has no edges to targets (or drugs). BLM-NII[32] is an improved version of BLM by integrating the procedure of neighbor-based interaction-profile inferring into the BLM method, which can find the targets for new drug candidates and identify the targeting drugs for new target candidates. However, too much emphasis on neighbor of BLM-NII tends to eliminate the local characteristics of each drug and target, maybe causing deterioration of prediction performance[32]. DT-Hybrid[33] is another improved version of BLM by adding domain-dependent biological knowledge through a similarity matrix. NBI[35] uses drug-target bipartite network topology similarity to infer new targets for known drugs. However, NBI cannot predict targets of a new drug (or drugs of a new target), because a new drug/target has no edges to targets/drugs, by which we cannot compute the connection scores. WNBI[46] is an improved version of NBI by considering the potency of binding affinity (or inhibitory activity) of the physical interactions between the chemical node and protein node to weight the edges among chemicals and proteins, or by using a new expression of initial resource distribution and taking into account the influence of resources associated with the receiver nodes to weight the nodes. WNBI is still cannot predict targets of a new drug (or drugs of a new target), but this bottleneck can be resolved by integrating the WNBI (or NBI) and DBSI (drug-based similarity inference) methods[35]. NetCBP[43] is a semi-supervised method, which uses the consistency in networks to measure whether the query drug and a target protein show coherent interaction with the known drug-target interactions. Although NetCBP shows the encouraging improvement, it depends heavily on the drug/ target similarity values. NRWRH[42] integrates three different networks (protein similarity network, drug similarity network and known drug-target interaction network) into a heterogeneous network to predict the drug-target associations by implementing the random walk on this heterogeneous network. This methodology shows excellent

performance in predicting new interactions. However, the drug-target interaction network is sparse because of the rare known drug-target interactions. Randomly walking on this sparse network, NRWRH maybe generates local solution.

Taken together, the limitations of current methods for predicting drug-target interactions are mainly in the following four aspects: i) Some methods randomly select the unknown drug-target interactions as negative samples, while most of these unknown drug-target interactions have not verified by biological experiments. ii) The performance of combing several distinct classifiers to obtain the final results is not strong enough. iii) Some methods cannot predict the potential target proteins for new drugs without any known target interaction information. iv) Some methods just utilize the drug's structural similarity and target protein sequence similarity.

To further improve the prediction accuracy of drug-target interactions and avoid the local solution of NRWRH, in this present article, we propose a new method (called LPMIHN) to infer the potential drug-target interactions by performing label propagation with mutual interaction information derived from heterogeneous network. LPMIHN consists of the following four steps. Firstly, constructing two similarity matrices of drug and target by fusing the drug's chemical similarity with drug topological similarity of drug-target interaction network, fusing the target protein sequence similarity with target topological similarity of drug-target interaction network, respectively; Secondly, establishing three networks (drug similarity network, target similarity network and known drug-target interaction bipartite network) to form a heterogeneous network; Thirdly, implementing label propagation on drug (or target) similarity sub-network to obtain the drug (or target) label network; Fourthly, implementing label propagation on the target (or drug) similarity sub-network, whose initial label information derived from the drug (or target) label network and the drug-target bipartite network; Finally, the most probable targets (or drugs) are selected according to the stable label scores of the walk.

LPMIHN is mainly different from NRWRH in three aspects. One is that drug/ target similarity network integrates the topological information of known drug-target interaction network. Another is that the label propagation (or random walk) is

implemented on the drug and target similarity networks, respectively. Thirdly, the initial label information of target/drug network comes from drug/target label network and known drug-target bipartite network.

Through extensive simulations on four benchmark datasets and two quantitative kinase bioactivity datasets, LPMIHN shows better performance than the existing state-of-the-art methods, such as BLM-NII, NetCBP and NRWRH. Furthermore, some new predicted drug-target interactions ranked in top were reported by publicly accessible datasets. It is anticipated that our LPMIHN algorithm can help us to find new or potential drug-target interactions, and provide useful information for drug design.

2 Materials

To facilitate benchmarking comparison with other state-of-art methods, we used the four drug-target interaction datasets from humans, namely enzymes (Es), ion channels (ICs), G-protein coupled receptors (GPCRs) and nuclear receptors (NRs), which were originally provided by Yamanishi *et al*.[40], and widely used as the benchmark binary interaction datasets of compounds targeting pharmaceutically useful target proteins [29, 31, 34, 35, 42-44, 47, 48]. These datasets are available at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Es dataset includes 445 drugs, 664 targets and 2926 known drug-target interactions. ICs dataset includes 210 drugs, 204 targets and 1476 known drug-target interactions. GPCRs dataset includes 223 drugs, 95 targets and 635 known drug-target interactions. NRs dataset includes 54 drugs, 26 targets and 90 known drug-target interactions.

Due to binary interaction datasets ignore many important characteristics of the drug-target interaction, such as dose-dependence and quantitative affinity, we use the same cutoff thresholds of $K_d \leq 30.00$ nM and $K_i < 28.18$ nM as Ref.[18] to binarize two large-scale quantitative kinase bioactivity datasets, i.e., kinase disassociation constant (K_d) dataset and kinase inhibition constant (K_i) dataset[49, 50], forming two binary interaction datasets which include 68 drugs, 442 targets and 1527 drug-target interactions for K_d dataset, and 1421 drugs, 156 targets and 3200 drug-target

interactions for K_i dataset. These two datasets are applied to evaluate the performance of our LPMIHN algorithm. The smaller is the K_d/K_i bioactivity, the higher is the interaction affinity between the chemical compound and the protein kinase.

Table 1 lists some statistics of each dataset including the total number of drugs (N_d), the total number of targets (N_t), the total number of interaction edges (E_{dt}), the total number of drugs that have only one targeting protein ($k_d(1)$), the total number of targets that have only one associated drug ($k_t(1)$), the average number of targets for each drug (avg. N_d), the average number of drugs for each target (avg. N_t), and the sparsity defined as that the total number of connected edges in real network is divided by the total number of linked edges in the complete graph.

Table 1 Statistical characteristics of six drug-target interaction datasets

Dataset	N_d	N_t	E_{dt}	$k_d(1)$	$k_t(1)$	avg. N_d	avg. N_t	Sparsity
Es	445	664	2926	177	288	6.58	4.41	0.0099
ICs	210	204	1476	81	23	7.03	7.24	0.0344
GPCRs	223	95	635	106	34	2.85	6.68	0.0299
NRs	54	26	90	39	8	1.67	3.46	0.0641
K_d	68	442	1527	4	97	22.46	3.45	0.0508
K_i	1421	156	3200	204	11	2.25	20.51	0.0144

3 Methods

Our LPMIHN method can be divided into two parts: constructing the heterogeneous network and separately implementing label propagation on the drug/target similarity networks.

3.1 Heterogeneous network

The heterogeneous network of drug-target interactions is composed of three typical networks of drug similarity network，target similarity network and known drug-target interaction bipartite graph network (see Fig. 1)

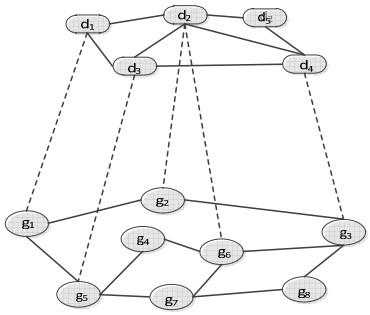


Figure 1 Drug-target interaction heterogeneous network model. Upper sub-network is the drug similarity network, underlying sub-network is the target protein similarity network and the intermediate layer is a drug-target interaction bipartite graph network.

The matrix S_d corresponding to the drug similarity network is composed of the chemical structure similarity matrix S_d^c and drug-target interaction profile-based drug similarity matrix S_d^{IP}. The matrix S_g corresponding to the target protein similarity network is composed of the protein sequence similarity matrix S_g^s and drug-target interaction profile-based target similarity matrix S_g^{IP}. The drug-target interaction adjacent matrix A corresponding to the drug-target interaction network is derived from the KEGG BRITE[2], BRENDA[51], SuperTarget [52] and DrugBank [53] databases. Here, S_d^c represents the chemical space, which was constructed by inferring the chemical structure similarity between drugs with SIMCOMP tool [54] based on the information obtained from the KEGG DRUG and KEGG LIGAND databases[2], and S_g^s represents the genomic space, which was created by calculating the normalized Smith-Waterman score [55] between any two amino acid sequences of target proteins. According to the study of literature [48] and [32], we can use the following formulas to calculate matrices S_d^{IP} and S_g^{IP}.

$$S_d^{IP}(d_i, d_j) = \exp(-\gamma_d \left\| x_i - x_j \right\|^2) \quad (1)$$

$$S_g^{IP}(g_i, g_j) = \exp(-\gamma_g \left\| y_i - y_j \right\|^2) \quad (2)$$

where, x_i is the interaction profile of drug d_i, y_i is the interaction profile of target g_i. The parameters γ_d and γ_g control the kernel bandwidth, which are set same as literature [48] and [32] in this paper. That is,

$$\gamma_d = 1/(\frac{1}{n}\sum_{i=1}^{n}|x_i|^2) \quad (3)$$

$$\gamma_g = 1/(\frac{1}{m}\sum_{i=1}^{m}|y_i|^2) \quad (4)$$

here, n and m are the total number of drugs and targets, respectively.

By integrating the interaction profile (i.e., drug-target interaction) information with the drug chemical and target genomic information, S_d and S_g are respectively defined as:

$$S_d = \lambda S_d^c + (1-\lambda) S_d^{IP} \quad (5)$$

$$S_g = \lambda S_g^s + (1-\lambda) S_g^{IP} \quad (6)$$

where parameter λ is the combination weight. Although Kronecker product is often used to combine two kinds of kernel matrices or similarity matrices, which is more sophisticated, the linear combination of two typical matrices can give the comparable performance with much lower computational complexity [48].

3.2 LPMIHN algorithm

Graph-based label propagation algorithms (LP) are the semi-supervised methods, which use the global network structure to improve the performance of classification and ranking [56-58]. Label propagation (LP) algorithm is closely related to the random walk (RW) algorithm. There are two major differences between LP and RW: i) LP fixes the labeled points: ii) the solution of LP is an equilibrium state while RW is dynamic w.r.t a time parameter t [56]. LP algorithm is mathematically identical to random walk with restart (RWR) algorithm if the similarity matrix W in RWR is normalized as $S = D^{-1} * W$, where D is the diagonal degree matrix with $D_{ii} = \sum_j W_{ij}$. In LP algorithms, some vertices (labeled data) are initialized with labels (e.g., 1) and other vertices (unlabeled data) are initialized with 0. The known label information on the vertices is iteratively propagated between the neighboring vertices and the propagation process will finally converge toward the unique global optimum by minimizing the quadratic criteria[59] . Until now, most graph-based label propagation

algorithms propagate label information on the single network or homo-network[58-60], which are not suitable for propagating label information across several sub-networks with different types of vertices and edges, because label propagation associated with the graph Laplacian of a signal network through a regularization framework ignores the difference among the sub-networks in a heterogeneous network. Thus, a heterogeneous label propagation (MINProp) algorithm was proposed by Hwang et al.[61] for discovering the disease gene on the disease-phenotype heterogeneous similarity network, which is superior to the original label propagation algorithm on a single network. In this paper, based on the assumption that similar drugs tend to target similar proteins and the framework of label propagation[56, 57], we will introduce a new label propagation algorithm, called LPMIHN, to infer the potential drug-target interactions. LPMIHN algorithm can be described in detail as follows:

Step1. For each drug-target dataset, we compute its drug/target similarity matrices S_d , S_t , and obtain its drug-target interaction adjacent matrix A to construct the heterogeneous network. Use the following equations to normalize the drug similarity matrix S_d, target similarity matrix S_t and drug-target interaction adjacent matrix A.

$$D = W_d^{-\frac{1}{2}} S_d W_d^{-\frac{1}{2}} \tag{7}$$

$$G = W_g^{-\frac{1}{2}} S_g W_g^{-\frac{1}{2}} \tag{8}$$

$$\tilde{A} = [\tilde{A}_{ij}], \qquad \tilde{A}_{ij} = \frac{A_{ij}}{\sum_j A_{ij}} \tag{9}$$

where W_d and W_g are the diagonal matrices whose diagonal elements are $W_{d,ii} = \sum_j S_{d,ij}$, $W_{g,ii} = \sum_j S_{g,ij}$ respectively, A_{ij} is the element of i row and j column in matrix A.

Step2. For a given query drug, we first implement label propagation on the drug similarity network by optimizing the following objective function to obtain state drug label network.

$$\min_f \left(\sum_{i,j} D_{ij}(f_i - f_j)^2 + \frac{1-\alpha}{\alpha}\sum_i (f_i - f_i^0)^2\right) \tag{10}$$

where f_i is the current label confidence score of drug d_i; f_i^0 is the initial label score of drug d_i; diffusion parameter α specifies the relative amount of the information from the initial label information to its neighbors. The solution of objective function Eq.10 can be obtained by iteratively performing the following equation.

$$f^t = \alpha D f^{t-1} + (1-\alpha) f^0, \alpha \in (0,1) \tag{11}$$

where, f^t is a n dimensional vector in which the ith element represents the label confidence score of drug d_i at time step t, and f^0 is a n dimensional initial label vector which derives from the known drug-target interaction network.

Step3. Implement label propagation on the target similarity network to predict which targets associate with the query drug by optimizing the following objective function.

$$\min_p \left(\sum_{i,j} G_{ij}(p_i - p_j)^2 + \frac{1-\alpha}{\alpha}\sum_i (p_i - p_i^0)^2\right) \tag{12}$$

where p_i is the current label confidence score of target g_i; p_i^0 is the initial label score of target g_i; diffusion parameter α specifies the relative amount of the information from the initial label information to its neighbors. The solution of objective function Eq.12 can be obtained by iteratively performing the following equation.

$$p^t = \alpha G p^{t-1} + (1-\alpha) p^0, \alpha \in (0,1) \tag{13}$$

where p^t is a m dimensional vector in which the ith element represents the label confidence score of target g_i at time step t; p^0 is a m dimensional initial label vector which is defined as follows.

$$p^0 = \frac{1-2\alpha}{\alpha} g^0 + \frac{\alpha}{1-\alpha} \tilde{A} f \tag{14}$$

where g^0 is a m dimensional initial label vector which derives from original target-drug interaction network, and f is the current label confidence score vector of drugs in the convergent drug label network. The first term is the initial label information of the targets, and the second term is the mutual interaction information which can capture the bi-cluster structures between the vertices in each pair of the sub-networks[61].

Step4. The vector p^t converges to its limit p^* after $\left\| p^t - p^{t-1} \right\| < \sigma$, in general, $\sigma = 10^{-9}$, and p^*gives the ranking score of every target for a query drug. Targets with maximum in p^* are considered as the most probable target of query drug.

For a given query target, we first implement label propagation on the target similarity network same as step3. Then, implement label propagation on the drug similarity network same as step3-4, give the ranking confidence scores of every drug for a query target.

4 Results and Discussion

4.1 Performance Evaluations

In order to illustrate the effectiveness of the strategies used in our LPMIHN algorithm, that is, the strategy of label propagation with mutual interaction information derived from heterogeneous network and the strategy of integrating the similarity information of drug and target with the topology information of drug-target interaction network, we introduce another two algorithms (MINProp and LPMIHN-S) to predict the potential drug-target interactions. MINProp algorithm proposed by Hwang[61] sequentially propagates the label information on S_d and S_g sub-network with the current label information derived from the known drug-target network and repeat this step until convergence. LPMIHN-S performs the label propagation on S_d^s and S_g^s sub-networks. The difference between LPMIHN and MINProp is the label propagating pathways. LPMIHN algorithm firstly propagates the label on the S_d sub-network, and then performs the label propagation on the S_g sub-network, but MINProp sequentially propagates the label on the S_d and S_g sub-networks. The only difference between LPMIHN and LPMIHN-S is that LPMIHN-S doesn't integrate the topology information of the known drug-target interactions.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, *K*-fold (e.g. 5-fold, 10-fold) crossover or subsampling test, and jackknife test (also called as leave-one-out test)[62-64]. In the three test methods, the jackknife

test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [65]. Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors [38, 63, 64, 66]. During the jackknife test, each known drug-target interaction in the dataset is singled out in turn as a test sample, and the remaining known drug-target interactions are used as training samples. However, jackknife test is more sensitive to over fitting than *K*-fold validation test, and independent dataset test is often used to evaluate the generalization ability of the predictors [65]. Therefore, we used the three cross-validation methods of jackknife, *k*-fold validation and independent tests to evaluate the performance of LPMIHN and other methods in the following experiments. The metrics of AUC and AUPR were used to measure the quality of the predicted drug-target interactions. AUC is the area under the receiver operating characteristic (ROC) curve which plots the true-positive rate (sensitivity) versus false-positive rate (1-specificity) at different cutoffs. AUPR is the area under the precision-recall curve which plots the ratio of true positives among all positive predictions for each given recall rate. For the prediction of drug-target interaction, AUPR is a more significant quality measure than AUC, as it punishes much more the existence of false positive drug-target interactions found among the best ranked prediction scores[28, 67] .

Table 2 gives the AUC and AUPR scores of the LPMIHN, MINProp and LPMIHN-S methods for the four binary drug-target interaction datasets of Enzymes (Es), Ion Channels (ICs), G-protein coupled receptors (GPCRs) and nuclear receptors (NRs), and two quantitative drug-target bioactivity datasets of the kinase disassociation constant (K_d) and the kinase inhibition constant (K_i). From table 2, we can see that the AUC scores of our LPMIHN method is little higher than that of MINProp and LPMIHN-S methods, but the AUPR scores is more high than that of MINProp and LPMIHN-S methods. The AUPR scores of our LPMIHN method are 0.9290, 0.9611, 0.9733, 0.9703, 0.7096 and 0.9944 on the enzyme, Ion Channel, G-protein coupled receptor, nuclear receptor, kinase disassociation constant and kinase inhibition constant datasets, respectively, which is 0.0799, 0.1197, 0.0463,

0.0319, 0.0729 and 0.1703 higher than that of MINProp method on the enzyme, Ion Channel, G-protein coupled receptor, nuclear receptor, kinase disassociation constant and kinase inhibition constant datasets, respectively, meaning that our label propagation strategy is superior to the strategy used in MINProp. The AUPR scores of our LPMIHN is also 0.1, 0.0696, 0.0232, 0.0476, 0.0411 and 0.0046 higher than that of LPMIHN-S method on the enzyme, Ion Channel, G-protein coupled receptor, nuclear receptor, kinase disassociation constant and kinase inhibition constant datasets, respectively, meaning that incorporating the network-based similarity can indeed improve the performance of LPMIHN. Especially, the efficacy of our label propagation strategy is better than that of the strategy of incorporating the network-based similarity.

Table 2 Performance of MINProp, LPMIHN-S and LPMIHN in jackknife (leave-one-out) test

Dataset	Methods	AUC	AUPR
Es	MINProp	0.9899	0.8491
	LPMIHN-S	0.9592	0.8290
	LPMIHN	0.9989	0.9290
ICs	MINProp	0.9837	0.8414
	LPMIHN-S	0.9917	0.8915
	LPMIHN	0.9985	0.9611
GPCRs	MINProp	0.9869	0.9370
	LPMIHN-S	0.9894	0.9501
	LPMIHN	0.9986	0.9733
NRs	MINProp	0.9729	0.9384
	LPMIHN-S	0.9804	0.9227
	LPMIHN	0.9960	0.9703
K_d	MINProp	0.9773	0.6367
	LPMIHN-S	0.9603	0.6685
	LPMIHN	0.9819	0.7096
K_i	MINProp	0.7375	0.8241
	LPMIHN-S	0.9985	0.9898
	LPMIHN	0.9995	0.9944

In order to investigate the potential false risk of only using the protein sequence similarity information, we just use one kind of similarity information (e.g., protein sequence similarity, GO similarity, protein-protein interaction network topological similarity, drug-target interaction profile similarity) to construct the target similarity

matrix S_g, then perform Label propagation on these target similarity networks and drug similarity network. For convenience, we name these methods as LPMIHN-SS, LPMIHN-GO, LPMIHN-PPI, LPMIHN-IP, respectively. The GO-based similarity matrix can be produced by computing the overlap rate of gene ontology annotations between any two proteins [23]. The PPI-based similarity matrix can be generated by computing the shortest distance between two target proteins in a human protein-protein interaction (PPI) network[68], and transforming these shortest distances into the similarity measures with Eq.1 in literature [69]. The sequence-based similarity and the drug-target interaction profile-based protein similarity matrices are the S_g^s and S_g^{IP}, respectively. The results of LPMIHN-SS, LPMIHN-GO, LPMIHN-PPI and LPMIHN-IP on the four benchmark datasets with jackknife test and 10CV test are shown the table 3 and table S1 (see supplementary). From table 3 and S1, we can see that the false positive rate (FPR) of four kinds of information sources is almost at the same level, the results of drug-target interaction profile information are the best in the four kinds of information sources, which indicate that the known drug-target interaction profile information can be used to improve the prediction performance, and the FPR of our LPMIHN method is not sensitive to the protein sequence similarity information.

Table 3 Results of LPMIHN-SS, LPMIHN-GO, LPMIHN-PPI and LPMIHN-IP on the four benchmark datasets with Jackknife test

Dataset	AUC				AUPR				FPR (%)			
	SS	GO	PPI	IP	SS	GO	PPI	IP	SS	GO	PPI	IP
Es	0.9593	0.9293	0.9543	0.99	0.884	0.6011	0.57	0.911	0.2	0.44	0.5	0.14
ICs	0.9918	0.9799	0.9915	0.9928	0.8912	0.8657	0.9195	0.9514	0.55	0.71	0.54	0.28
GPCRs	0.9894	0.9792	0.9844	0.9976	0.9508	0.9259	0.9469	0.9623	0.25	0.32	0.3	0.18
NRs	0.9937	0.9613	0.9684	0.9956	0.9579	0.8961	0.9155	0.9669	0.53	0.91	0.76	0.53

Notes: SS, GO, PPI and IP represent the LPMIHN-SS, LPMIHN-GO, LPMIHN-PPI and LPMIHN-IP, respectively.

4.2 Compared with other state-of-the-art methods

We compared the performance of LPMIHN with two recent state-of-the-art network methods of BLM-NII[32] and NRWRH[42] on the binary drug-target interaction datasets of Enzymes (Es), Ion Channels (ICs), G-protein coupled receptors (GPCRs), nuclear receptors (NRs), and two quantitative kinase bioactivity datasets of the kinase disassociation constant (K_d) and the kinase inhibition constant (K_i) in jackknife test. The results of BLM-NII, NRWRH and LPMIHN are shown in table 4, from which we can see that the performance of our LPMIHN is superior to other two methods both in terms of AUC and AUPR for all the six datasets. The AUC scores of LPMIHN is 0.008~0.056 higher than that of BLM-NII, and the AUPR scores of LPMIHN is 0.01~0.1735 much higher than that of BLM-NII on the Ion channels, G-protein coupled receptors , nuclear receptors datasets and two quantitative kinase bioactivity datasets of K_d, K_i. In addition, the AUPR is a more significant quality measure than AUC [28, 67]. The reason that AUPR scores both of LPMIHN and BLM-NII are equal on the Enzymes dataset may be that the network constructed with enzymes dataset is too sparse, whose sparsity is just 0.0099. These results show that our LPMIHN can effectively predict the drug-target interactions, especially for dense interaction network.

Table 4 Comparison with existing network approaches on binary drug-target interaction datasets in jackknife test

Dataset	Method	AUC	AUPR
Es	BLM-NII	0.9880	0.9290
	NRWRH	0.9533	0.6338
	LPMIHN	0.9989	0.9290
ICs	BLM-NII	0.9900	0.9500
	NRWRH	0.9707	0.5908
	LPMIHN	0.9985	0.9611
GPCRs	BLM-NII	0.9840	0.8650
	NRWRH	0.9447	0.6739
	LPMIHN	0.9986	0.9733
NRs	BLM-NII	0.9810	0.8660
	NRWRH	0.8665	0.6630
	LPMIHN	0.9960	0.9703
K_i	BLM-NII	0.9820	0.8209
	NRWRH	0.8213	0.1602
	LPMIHN	0.9995	0.9944
K_d	BLM-NII	0.9252	0.6270
	NRWRH	0.8601	0.2484
	LPMIHN	0.9818	0.7096

To test the robustness of the presented approach, we also performed 10/5-fold cross-validation (10CV, 5CV) test and compared with the reported results of other methods: NBI[35], DNBSI[35], TBSI[35], BLM-NII[47], NetLapRLS[29], KBMF2K[34], NetCBP[43] and NRWRH[42]. The AUC values of these methods on the four benchmark datasets in 10/5CV test can be found in tables S2-S5 of the Supplementary materials. From tables S2-S5, we can see that the AUC value of LPMIHN is still higher than that of other methods in 10CV (or 5CV) test, which indicates that the robustness of our LPMIHN is better.

In order to test the generalization ability of our LPMIHN, we used one of the two external validation datasets in ref.[27] Table S6 collected from DrugBank and KEGG databases to evaluate the performance of our LPMIHN. The external dataset contains 86 approved and investigated drugs and 44 GPCRs, and there are 256 known interaction relationships among these drugs and GPCRs. We removed these drug-target interactions from GPCRS benchmark dataset, and the remaining drug-target interactions are used as training samples to train NRWRH and LPMIHN models. The results of NRWRH and LPMIHN on this external dataset are shown in table 5, from which we can see that LPMIHN achieved higher performance in terms of AUC and AUPR, indicating that our LPMIHN method has a better generalization performance for predicting drug-target interactions.

Table 5 Results of NRWRH and LPMIHN in independent dataset test

Methods	AUC	AUPR
NRWRH	0.818	0.3567
LPMIHN	0.9749	0.8885

To further evaluate the generalization of our LPMIHN algorithm, we randomly select 10% drugs from K_d and K_i databases respectively as the testing subsets, and **all known interaction relationships with targets of these test drugs are deleted**. The residual 90% drugs are as the training subsets to train the LPMIHN. The average AUC values of 5 independent tests arrive at 0.761 and 0.8094 for K_d and K_i respectively,

which indicates that our LPMIHN algorithm has a certain extent of over-fitting risk, but it has the better generalization.

To further confirm the superior performance of LPMIHN, considering the fact that high-confidence prediction results are interested by the drug developers, we also compared the results of BLM-NII, NRWRH with LPMIHN on the sensitivity, positive predictive value (PPV) and Matthews correlation coefficient (MCC) when the upper 1% and 3% target in the prediction list is chosen as a threshold. Table 5 shows the comparative results of BLM-NII, NRWRH with LPMIHN for NR dataset.

Table 6 Performance comparison of BLM-NII, NRWRH and LPMIHN on top 1% and 3% ranked in the prediction list

	Method	Top 1%			Top 3%		
		Sensitivity	PPV	MCC	Sensitivity	PPV	MCC
λ=0.5	BLM-NII	0.156	1.0	0.383	0.456	0.976	0.654
	NRWRH	0.156	1.0	0.383	0.378	0.810	0.534
	LPMIHN	0.156	1.0	0.383	0.467	1.00	0.671

4.3 Effect of the parameters

There are two parameters λ and α in our LPMIHN algorithm. λ is a combination weight parameter, which controls the contribution of two kinds of information of the drug chemical (or target sequence) similarity and the network-based drug/target similarity. By selecting different λ values (varying from 0 to 1 with scale 0.05) to simulate, the AUC and AUPR values for Nuclear receptors dataset are shown in figure 2, in which we found that the AUC value is almost equal in the range $0.05 \le \lambda \le 0.5$, and decreased slowly in the range $0.5 < \lambda \le 0.95$; AUPR value kept almost constant value in the range $0.05 < \lambda \le 0.35$, and decreased slowly in the range $0.35 < \lambda < 0.95$. But for value $\lambda = 0$, and $\lambda = 1$, the AUC value and AUPR value are all very small. In this work, we selected $\lambda = 0.5$.

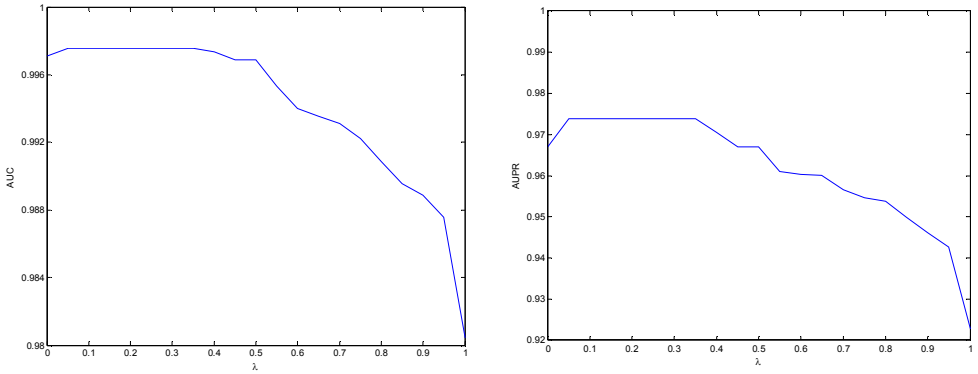


Figure 2 The relationship between the parameter λ and AUC, AUPR for LPMIHN algorithm on the NR dataset

α is a diffusion parameter, which adjusts the relative amount of the information from the initial label information to its neighbors. By selecting different α values (varying from 0.05 to 0.95 with scale 0.05) to simulate, the AUC and AUPR results for Nuclear receptors dataset are shown in figure 3, in which we found that the AUC value is almost equal in the range $\alpha \le 0.4$, decreased quickly in the range $0.4 < \alpha \le 0.5$, decreased gradually in the range $0.5 < \alpha \le 0.7$, and decreased quickly in the range $\alpha > 0.7$. In this work, we fixed $\alpha = 0.2$. The parameters effects of λ and α for Enzymes, ICs, GPCRs, K_d and K_i datasets are shown in supplement figures S1-S4 of supplementary materials.

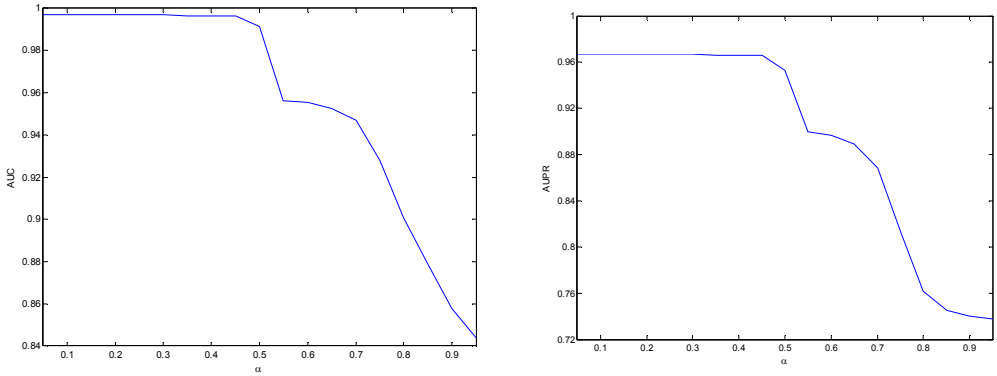


Figure 3 The relationship between the parameter α and AUC, AUPR for LPMIHN algorithm on the NR dataset

4.4 Analysis of LPMIHN computational complexity

The computational complexity of our LPMIHN algorithm is comprised of three parts: separately performing label propagation on drug target similarity networks, updating initial label. In the phrase of separately performing label propagation on drug and target similarity network, the time complexity is in the order of $O(n^2)$ and $O(m^2)$, where n and m are the total number of nodes in drug and target similarity networks, respectively. If separately iteratively performing k_1 *and* k_2 times label propagation on the drug and target similarity networks to obtain the final label confidence scores of all the vertices, then, the total time complexity of performing label propagation on drug and target similarity networks is in the order of $O(k_1*n^2 + k_2*m^2)$. For the phase of updating initial label, it should perform label propagation on the drug-target interaction bipartite network, thus, thus the maximum time complexity is in the order of $O(n*m)$. Therefore, the overall time complexity of LPNMIH algorithm is in the order of $O(n*m+ k_1*n^2 + k_2*m^2)$. While the time complexity of NRWRH is in the order of $O(k*(n+m)^2)$, where k is iterative times of random walk. Obviously, the time complexity of NRWRH is smaller than that of NRWRH.

4.5 Analysis of the predicting potential drug-target interactions

Except for predicting known interactions, detecting potential unknown drug-target interactions is more significant. In order to illustrate the performance of our LPMIHN algorithm in predicting unknown drug-target interactions, we used all the known drug-target interactions on the four benchmark datasets as the initial labels, and performed LPMINH algorithm to rank the unknown drug-target interactions. According to their ranking scores, we extracted the top 50 potential drug-target interactions for each of the four datasets, and manually checked these predicted interactions in the latest online versions of ChEMBL[5], KEGG DRUG[6], Drugbank[8] and SuperTarget[7] datasets. Here, we reported only the top five potential interactions for each dataset and other potential interactions are listed in supplementary Tables S6-S9, in which we found that some target proteins target more drugs. For example, within the top 50 predicted drug-target interactions, hsa590 is the

potential target of 14 drugs in Enzymes dataset; hsa3781 and hsa6531 is the potential target of 6 drugs in Icon Channel dataset; hsa152 is the potential target of 9 drugs in GPCR dataset; hsa2100 is the potential target of 11 drugs in Nuclear Receptor dataset.

Table 7 Top five predicted potential drug-target interactions on the four benchmark datasets

Dataset	Rank	Pair	Annotation	Source
Es	1	D00097 hsa:5743	Salicylic acid (JP16/USP) prostaglandin-endoperoxide synthase 2	ChEMBL, Drugbank
	2	D00449 hsa:5742	Sulfinpyrazone (JAN/USP/INN) Prostaglandin-endoperoxide synthase 1	ChEMBL
	3	D00947 hsa:4129	Linezolid (JAN/USAN/INN) monoamine oxidase B	ChEMBL
	4	D05458 hsa:4128	Phentermine (USAN/INN Amine oxidase [flavin-containing] A	Drugbank
	5	D00005 hsa:4128	Flavin adenine dinucleotide(JAN) Amine oxidase [flavin-containing] A	Drugbank
ICs	1	D05453 hsa:3738	Phencyclidine hydrochloride (USAN) potassium voltage-gated channel, shaker-related subfamily, member 3	
	2	D02207 hsa:3782	Tubocurarine chloride (USP) Potassium intermediate/small conductance calcium-activated channel subfamily N member 3	
	3	D02356 hsa:6323	Verapamil (USAN/INN) Voltage-gated sodium channel type I alpha	
	4	D00616 hsa:9312	Diltiazem hydrochloride(JP16/USP) Potassium voltage-gated channel Shab-related subfamily B, member 2	
	5	D03830 hsa:9312	Diltiazem malate (USAN) Potassium voltage-gated channel Shab-related subfamily B, member 2	
GPCRs	1	D02910 hsa:154	Amiodarone (USAN/INN) adrenoceptor beta 2, surface	
	2	D00454 hsa:152	Olanzapine (JAN/USAN/INN) adrenoceptor alpha 2C	Drugbank
	3	D02358 hsa:154	Metoprolol (USAN/INN) adrenoceptor beta 2, surface	Drugbank ChEMBL
	4	D04625 hsa:154	Isoetharine (USP) adrenoceptor beta 2, surface	KEGG ChEMBL
	5	D00283 hsa:152	Clozapine(JAN/USP/INN) adrenoceptor alpha 2C	Drugbank
NRs	1	D00585 Has:2099	Mifepristone (USAN/INN) estrogen receptor 1	ChEMBL
	2	D00182 Has:2099	Norethisterone (JP16) estrogen receptor 1	ChEMBL
	3	D00951 Has:2099	Medroxyprogesterone acetate (JAN/USP) estrogen receptor 1	Drugbank
	4	D00690 Has:2099	Mometasone furoate (JAN/USP) estrogen receptor 1	
	5	D01217 Has:2099	Dydrogesterone (JP16/USP/INN) estrogen receptor 1	

Table 7 lists the top 5 predicted potential interactions for each dataset, in which we can see that top five predicted drug-target interactions on Enzymes dataset, and

four of the five predicted interactions on GPCR and three of the five on Nuclear receptor datasets are reported in ChEMBL, KEGG DRUG, Drugbank and SuperTarget databases. Other predicted interactions that are not reported yet may also be verified by the experiments in the future. These results show that our LPMIHN algorithm can effectively mine the unknown drug-target interactions.

5 Conclusions

In this work, LPMIHN was developed to predict the potential drug-target interactions by integrating multi-source information to construct a heterogeneous network, and introducing a new strategy of network-based label propagation with mutual interaction information derived from heterogeneous network. The originality of LPMIHN mainly lies in that it separately performs label propagation on drug/target similarity networks, while the initial label information of target (or drug) network comes from drug (or target) label network and known drug-target interaction bipartite network. The independent label propagation on each similarity network can explore the cluster structure in its network, and the label information from the other networks can be used to capture mutual interactions between the nodes in each pair of the similarity networks. We used four benchmark datasets of enzymes, ion channels, GPCRs and nuclear receptors, and two quantitative kinase bioactivity datasets of the kinase disassociation constant (K_d) and the kinase inhibition constant (K_i) to demonstrate the performance of LPMIHN algorithm in the jackknife, *K*-fold cross-validation and independent tests. The results show that the strategies of label propagation and information fusion are effective for predicting the drug-target interactions. Comparison with other recent state-of-the-art methods, such as NetlabRLS, KBMF2K, NBI, BLM-NII, NetCBP and NRWRH, our LPMIHN algorithm obtains the best results in terms of AUC and AUPR. We also predicted the unknown drug-target interaction on the four benchmark datasets, and found that many top predicted potential interactions are reported in the latest publicly available drug-target databases, meaning that LPMIHN can help drug developers to find new or potential drug-target interactions.

Despite the encouraging improvement, our method performs the label propagation in the drug similarity network and target similarity network, respectively. From a technical viewpoint, the performance of our method could be improved by sequentially propagating the label information in the drug similarity network, target similarity network and drug-target heterogeneous network. However, it does not get the better results by using this way to predict the drug-target interactions. This is due to the drug-target interaction bipartite network too sparse, and it will result in local solutions. Thus, it would be an interesting future work to explore the label propagation way on the sparse heterogeneous network.

Funding

This paper was supported by the National Natural Science Foundation of China (No. 91430111, 61473232, 61170134).

References

[1] C.M. Dobson, Chemical space and biology, Nature, 432 (2004) 824-828.
[2] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, Nucleic acids research, 34 (2006) D354-D357.
[3] B.R. Stockwell, Chemical genetics: ligand-based discovery of gene function, Nature Reviews Genetics, 1 (2000) 116-125.
[4] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, Database resources of the national center for biotechnology information, Nucleic acids research, 39 (2011) D38-D51.
[5] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic acids research, 40 (2012) D1100-D1107.
[6] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic acids research, (2011) gkr988.
[7] N. Hecker, J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, M.K. Gilson, P.E. Bourne, R. Preissner, SuperTarget goes quantitative: update on drug–target interactions, Nucleic acids research, (2011) gkr912.
[8] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, DrugBank 3.0: a comprehensive resource for â€˜omicsâ€ research on drugs, Nucleic acids research, 39 (2011) D1035-D1041.
[9] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework, Bioinformatics, 26 (2010) i246-i254.

[10] S. Whitebread, J. Hamon, D. Bojanic, L. Urban, Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development, Drug discovery today, 10 (2005) 1421-1433.
[11] S.J. Haggarty, K.M. Koeller, J.C. Wong, R.A. Butcher, S.L. Schreiber, Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays, Chemistry & biology, 10 (2003) 383-396.
[12] T. Klabunde, G. Hessler, Drug Design Strategies for Targeting G‐Protein‐Coupled Receptors, Chembiochem, 3 (2002) 928-944.
[13] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, Nat Biotechnol, 25 (2007) 71-75.
[14] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework, Bioinformatics, 26 (2010) i246-254.
[15] F. Cheng, W. Li, Z. Wu, X. Wang, C. Zhang, J. Li, G. Liu, Y. Tang, Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space, Journal of chemical information and modeling, 53 (2013) 753-762.
[16] F. Cheng, W. Li, Y. Zhou, J. Li, J. Shen, P.W. Lee, Y. Tang, Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (PTDMs), Molecular BioSystems, 9 (2013) 1316-1325.
[17] J. Li, Z. Wu, F. Cheng, W. Li, G. Liu, Y. Tang, Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology, Scientific reports, 4 (2014).
[18] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, T. Aittokallio, Toward more realistic drug–target interaction predictions, Briefings in bioinformatics, (2014) bbu010.
[19] J. Ballesteros, K. Palczewski, G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin, Current opinion in drug discovery & development, 4 (2001) 561.
[20] H. Ding, I. Takigawa, H. Mamitsuka, S. Zhu, Similarity-based machine learning methods for predicting drug–target interactions: a brief review, Briefings in Bioinformatics, 15 (2014) 734-747.
[21] H. Liu, J. Sun, J. Guan, J. Zheng, S. Zhou, Improving compound–protein interaction prediction by building up highly credible negative samples, Bioinformatics, 31 (2015) i221-i229.
[22] L. Nanni, A. Lumini, S. Brahnam, A set of descriptors for identifying the protein–drug interaction in cellular networking, Journal of theoretical biology, 359 (2014) 120-128.
[23] Z. Mousavian, A. Masoudi-Nejad, Drug-target interaction prediction via chemogenomic space: learning-based methods, Expert opinion on drug metabolism & toxicology, 10 (2014) 1273-1287.
[24] F. Cheng, Z. Zhao, Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, Journal of the American Medical Informatics Association, 21 (2014) e278-e286.
[25] Y. Tabei, Y. Yamanishi, Scalable prediction of compound-protein interactions using minwise hashing, BMC systems biology, 7 (2013) S3.
[26] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, Y. Yamanishi, Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers, Bioinformatics, 28 (2012) i487-i494.
[27] F. Cheng, Y. Zhou, J. Li, W. Li, G. Liu, Y. Tang, Prediction of chemical–protein interactions: multitarget-QSAR versus computational chemogenomic methods, Molecular BioSystems, 8 (2012) 2373-2384.

[28] T. van Laarhoven, S.B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics, 27 (2011) 3036-3043.
[29] Z. Xia, L.-Y. Wu, X. Zhou, S.T. Wong, Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces, BMC systems biology, 4 (2010) S6.
[30] Y.-C. Wang, Z.-X. Yang, Y. Wang, N.-Y. Deng, Computationally probing drug-protein interactions via support vector machine, Letters in Drug Design & Discovery, 7 (2010) 370-378.
[31] X. Zheng, H. Ding, H. Mamitsuka, S. Zhu, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1025-1033.
[32] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, J. Zheng, Drug–target interaction prediction by learning from local information and neighbors, Bioinformatics, 29 (2013) 238-245.
[33] S. Alaimo, A. Pulvirenti, R. Giugno, A. Ferro, Drug-target interaction prediction through domain-tuned network-based inference, Bioinformatics, 29 (2013) 2004-2008.
[34] M. Gönen, Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization, Bioinformatics, 28 (2012) 2304-2310.
[35] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference, Plos Computational Biology, 8 (2012) 357-372.
[36] Y.-C. Wang, C.-H. Zhang, N.-Y. Deng, Y. Wang, Kernel-based data fusion improves the drug–protein interaction prediction, Computational biology and chemistry, 35 (2011) 353-362.
[37] K. Bleakley, Y. Yamanishi, Supervised prediction of drug-target interactions using bipartite local models, Bioinformatics, 25 (2009) 2397-2403.
[38] S.-W. Zhang, W. Chen, F. Yang, Q. Pan, Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach, Amino Acids, 35 (2008) 591-598.
[39] L. Jacob, J.-P. Vert, Protein-ligand interaction prediction: an improved chemogenomics approach, Bioinformatics, 24 (2008) 2149-2156.
[40] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug–target interaction networks from the integration of chemical and genomic spaces, Bioinformatics, 24 (2008) i232-i240.
[41] L. Jacob, J.P. Vert, Protein-ligand interaction prediction: an improved chemogenomics approach, Bioinformatics, 24 (2008) 2149-2156.
[42] X. Chen, M.-X. Liu, G.-Y. Yan, Drug–target interaction prediction by random walk on the heterogeneous network, Molecular BioSystems, 8 (2012) 1970-1978.
[43] H. Chen, Z. Zhang, A semi-supervised method for drug-target interaction prediction with consistency in networks, PloS one, 8 (2013) e62975.
[44] K. Bleakley, Y. Yamanishi, Supervised prediction of drug–target interactions using bipartite local models, Bioinformatics, 25 (2009) 2397-2403.
[45] S. Alaimo, A. Pulvirenti, R. Giugno, A. Ferro, Drug–target interaction prediction through domain-tuned network-based inference, Bioinformatics, 29 (2013) 2004-2008.
[46] F. Cheng, Y. Zhou, W. Li, G. Liu, Y. Tang, Prediction of chemical-protein interactions network with weighted network-based inference method, (2012).
[47] J.P. Mei, C.K. Kwoh, P. Yang, X.L. Li, J. Zheng, Drug-target interaction prediction by learning from local information and neighbors, Bioinformatics, 29 (2013) 238-245.

[48] T. van Laarhoven, S.B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug-target interaction, Bioinformatics, 27 (2011) 3036-3043.
[49] J.T. Metz, E.F. Johnson, N.B. Soni, P.J. Merta, L. Kifle, P.J. Hajduk, Navigating the kinome, Nature chemical biology, 7 (2011) 200-202.
[50] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, P.P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, Nature biotechnology, 29 (2011) 1046-1051.
[51] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, D. Schomburg, BRENDA, the enzyme database: updates and major new developments, Nucleic acids research, 32 (2004) D431-D433.
[52] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E.G. Urdiales, A. Gewiess, L.J. Jensen, SuperTarget and Matador: resources for exploring drug-target relationships, Nucleic acids research, 36 (2008) D919-D922.
[53] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic acids research, 36 (2008) D901-D906.
[54] M. Hattori, Y. Okuno, S. Goto, M. Kanehisa, Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, Journal of the American Chemical Society, 125 (2003) 11853-11865.
[55] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, Journal of molecular biology, 147 (1981) 195-197.
[56] X. Zhu, Z. Ghahramani, Learning from labeled and unlabeled data with label propagation, Citeseer, 2002.
[57] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, Advances in neural information processing systems, 16 (2004) 321-328.
[58] Y. Bengio, O. Delalleau, N. Le Roux, Label propagation and quadratic criterion, Semi-supervised learning, 10 (2006).
[59] M.S.T. Jaakkola, M. Szummer, Partially labeled classification with Markov random walks, Advances in neural information processing systems (NIPS), 14 (2002) 945-952.
[60] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using gaussian fields and harmonic functions, ICML, 2003, pp. 912-919.
[61] T. Hwang, R. Kuang, A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery, SDM, SIAM, 2010, pp. 583-594.
[62] X.-N. Fan, S.-W. Zhang, lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning, Molecular BioSystems, (2015).
[63] S.-W. Zhang, Y.-F. Liu, Y. Yu, T.-H. Zhang, X.-N. Fan, MSLoc-DT: A new method for predicting the protein subcellular location of multispecies based on decision templates, Analytical biochemistry, 449 (2014) 164-171.
[64] S.-W. Zhang, L.-Y. Hao, T.-H. Zhang, Prediction of protein–protein interaction with pairwise kernel Support Vector Machine, International journal of molecular sciences, 15 (2014) 3220-3233.
[65] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, Journal of theoretical biology, 273 (2011) 236-247.
[66] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic acids research, (2013) gks1450.

[67] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233-240.
[68] T.K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, Human protein reference database—2009 update, Nucleic acids research, 37 (2009) D767-D772.
[69] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, R. Sharan, Combining drug and gene similarity measures for drug-target elucidation, Journal of computational biology, 18 (2011) 133-145.